TU Clausthal

Clausthal University of Technology

ICOLE 2009, Lessach, Austria

Jacek Blazewicz, Klaus Ecker, Barbara Hammer (Eds.)

IfI Technical Report Series IfI-09-12

Department of Informatics
Clausthal University of Technology

Impressum

Publisher: Institut für Informatik, Technische Universität Clausthal
Julius-Albert Str. 4, 38678 Clausthal-Zellerfeld, Germany
Editor of the series: Jürgen Dix
Technical editor: Michael Köster
Contact: michael.koester@tu-clausthal.de
URL: http://www.in.tu-clausthal.de/forschung/technical-reports/
ISSN: 1860-8477

The IfI Review Board

ICOLE - 2009
German - Polish Workshop on Computational Biology, Scheduling and Machine Learning
Lessach, 25.05. - 29.05.2009

Jacek Blazewicz, Klaus Ecker, Barbara Hammer (Eds.)

Managing Editor: Bassam Mokbel

Contents

Preview

***Jacek Blazewicz*[1,2], *Klaus Ecker*[3,4], *Barbara Hammer*[5,6]**

From May 24th to 30th, 2009, twenty-two scientists from Poznan University of Technology, Clausthal University of Technology, and University of Leipzig met in Lessach, Austria, to continue the tradition of Polish - German Workshops on Computational Biology, Scheduling, and Machine Learning. The aim was to present their current research, discuss scientific questions, and exchange their ideas. The seminar centered around topics in Scheduling, Bioinformatics, and Machine Learning, covering fundamental theoretical aspects as well as recent applications, partially in the frame of international European projects or innovative industrial cooperations. This volume contains a collection of extended abstracts which accompany these talks to give some insight into the research presented in Lessach.

The first six contributions in this volume center around scheduling problems. The adjacency of elegant mathematical modelling and innovative applications in scheduling is demonstrated in a contribution by J. Blazewicz et al. which models berth allocation for great harbors such as Hong Kong as a moldable task scheduling problem. Complex applications of scheduling in car manufactoring are presented in an approach by G. Pawlak et al. which proposes a model based on multi-agent systems particularly suited for make-to-order production. Similarly, a study of assembly line balancing to increase the production rate preseted by M. Morze et al. has direct applications to car manufactoring. Issues of implementation are highlighted in the context of revenue maximization for which different exact and heuristic algortihms are presented and implemented on parallel machines in a contribution by J. Juraszek et al. The relevance of scheduling for crisis management

[1]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[2]E-mail: jblazewicz@cs.put.poznan.pl
[3]Center for Intelligent, Distributed and Dependable Systems, Ohio University, Athens, USA
[4]E-mail: ecker@ohio.edu
[5]Department of Informatics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
[6]E-mail: hammer@in.tu-clausthal.de

is demonstrated in the frame of work for the European project SIMCA in a contribution by M. Tanas et al., where scheduling tasks to restore safety and security after crisis are modeled in terms of classical FMS problems. A more theoretical view is taken in the contribution of P. Gawron and R. Walkowiak where the bin packing problem, which offers a suitable formalization e.g. in the context of parallel process scheduling, is tackled and a scheme to arrive at exact solutions using dynamic programming is developed.

Problems in bioinformatics are the focus of the next six contributions in this volume. The problem to find functional regions in DNA sequences, more precisely to detect cis-regulatory modules is tackled within a flexible model accompanied by an efficient implementation and demonstration of the usability for the model plant *A. thaliana* in a contribution by K. Ecker. The same model plant is considered in a contribution by M. Milostan et al. which centers around protein structure prediction, in this case exemplary studies for dicer like proteins. Two further approaches address the problem of protein structures, one approach by M. Szachniuk et al. proposes an elegant graph structure which represents sequential resonance assignment in NMR spectroscopy recorded for RNA structure, while another contribution by P. Lukasiak et al. presents a web-based platform for protein domain identification and function prediction based on similarity search in data bases of known structures. The problem of appropriate data bases and corresponding analysis tools is addressed on a higher level in another contribution by P. Lukasiak et al.: in the frame of the CompuVac European project, frontier research in data bases and mining tools for next generation vaccines is performed. Concrete studies of the effectivity of treatment and its problems based on RNA information is presented in the context of the hepatitis c virus in a contribution by S. Wasik et al.

Because of the often only vaguely defined problems and noisy data sets, bioinformatics is often connected to machine learning or softcomputing tools which can help to get suitable results in these settings. Four contributions exemplarly consider machine learning and optimization heuristics in the context of bioinformatics problems. NP hard problems such as sequencing by hybridization can be efficiently tackled in the context of hyperheuristics as presented by A. Swiercz. One approach by H. Cwiek et al. explains the foundation of microarray experiments and typical problems and research questions which arise in the context of microarray expression profile analysis. Spectra such as NMR spectra or mass spectra constitute two further data forms which occur frequently in bioinformatics. In this context, an approach by F.-M. Schleif et al. presents metabolic profiling studies of stem cell extracts using 1-H NMR. Another approach presented by S. Simmuteit et al. focusses on mass spectra and the possibility to identify and deconvolute mass spectra using evolving trees and sparse coding, respectively.

Finally, four contributions center around core machine learning problems

and algorithmic development. The problem of determining relevant factors in generative data models based on auxiliary information is tackled in work presented by A. Gisbrecht in the frame of a new model for the popular generative topographic map. Similar deterministic methods in the supervised and unsupervised setting are revisited in another approach by B. Hammer, and it is demonstrated how matrix learning offers an improvement of the results as well as an interface to visualization models. The transfer of unsupervised visualization and mining tools to even more general data which are described by a dissimilarity matrix only and methods to arrive at efficient linear implementations are considered in a contribution by A. Hasenfuss and B. Hammer. Similarly, a contribution by T. Villmann and B. Hammer addresses general data in terms of kernels, and discusses the applicability to powerful supervised learning algorithms.

The workshop continued a series of similar events which took place in Daublebsky's wonderful house in Lessach and, as always, was a great success due to its international participants and the scientifically fruitful and cooperative atmosphere. Apart from the scientific merrits, this year's seminar came up with a few highlights which demonstrate the excellent possibilities offered by the house and its surroundings. We just mention two of them: A new record for climbing Gumma in 65 minutes without the help of additional oxygen or ethanol was established, and the unprecedented commitment of the Friday's cooking team contributed to one of the world's coolest barbecues. Our particular thanks for a perfect organization of the workshop go to Alexander Hasenfuss as spiritus movens of the seminar and Bassam Mokbel as managing editor of the abstract collection.

Lessach, May, 2009
Jacek Blazewicz, Klaus Ecker, Barbara Hammer

Berth allocation as a moldable task scheduling problem

***Jacek Blazewicz*[1,2], *T. C. Edwin Cheng*[3], *Maciej Machowiak*[2], *Cayda Oguz*[3,4]**

Acknowledgments: The work described in this paper was partially supported by a grant from the Research Grant Council of Hong Kong (Project No. PolyU 5193/01E) and the KBN grant (No. 4T11C03925).

In this paper, the allocation problem of berths to the incoming ships is modelled by moldable tasks scheduling problem. This model considers the tasks as the ships and the processors as quay cranes assigned to these ships. Since the duration of berthing for a ship depends on the number of quay cranes allocated to the ship, the use of moldable task scheduling model is substantiated. In the model, the processing speed of a task is considered to be a non-linear function of the number of processors allocated to it. A suboptimal algorithm, which starts from the continuous version of the problem (i.e. where the tasks may require a fractional part of the resources) to obtain a feasible solution to the discrete version of the problem, is presented. The computational experiments conducted showed that the suboptimal algorithm has a very good average behavior.

[1]E-mail: jblazewicz@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Departament of Logistics, The Hong Kong Polytechnic University, Hong Kong SAR
[4]Departament of Industrial Engineering, Koc University, Turkey

Multilayer agent system for scheduling and control in a car factory

***Grzegorz Pawlak*[2], *Slawomir Walkowski*[1,2], *Tomasz Zurkowski*[2]**

1 Introduction

Many car factories produce several car models at one time. Moreover, the equipment of produced cars is often very diverse. That is why some car factories use make-to-order (MTO) production system [3]. It means that every single car is manufactured as a unique one, from the planning stage until releasing the final product. Each car gets its own identification number. Production is planned as a sequence of these identifiers.

A factory which uses MTO system wants to preserve the sequence through all stages of production or at least to know how the sequence will look like in the certain points in the production lines. Unfortunately, this is a difficult task because of a complicated structure of some manufacture areas. It concerns particularly early stages of production, like car body shop, where production of each car is distributed among several parallel lines. They join in some places, including switches and buffers. At these positions the original sequence of cars is open to disruptions. Also other specific areas, like quality control or transport between factory plants, may affect the order of cars being produced.

2 Problem description

In the typical approach, production planning, controlling and monitoring are managed centrally. In most cases decisions about production are taken and implemented by a hierarchy of employers. It leads to situations when failures and anomalies cause the production to be stopped in whole areas of

[1]E-mail: swalkowski@skno.cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology

the factory. Moreover, it is extremely difficult to care about a sequence of cars manually. Thus, the sequence is likely to be disturbed. All these issues lead to decreasing the factory efficiency. We would like to solve this problem – preserve the sequence in the production lines and make the production itself more resistant to planned and unexpected changes.

Before working out a solution, we define some measures of the sequence quality. Simple conclusion that the order of cars has changed or not is not enough for us. We need a sequence evaluation method to determine to what extent the sequence has changed between points in the production lines (called status points).

One proposal of such quality index is PKG (from German *Perlenketteguete* – quality of a string of pearls). In this approach we select a range of P cars in a sequence in the first status point – we call it input window. Then, we must find out in which fragment of a sequence in the second status point this range should appear. There are several possibilities to determine it. The one which is most reasonable and which we will use in this paper is the following. We count the average number of cars which are in the production line between two status points. We call this number a circulation gap G. We wait until G cars pass the second status point. Then, we take first P cars which appear in the second point and treat it as the output window.

To calculate the PKG index we use the following formula:

$$PKG = \frac{P - P_{delayed} - P_{behind\ window}}{P}$$

where $P_{delayed}$ denotes the number of delayed cars in the output window and $P_{behind\ window}$ denotes the number of cars from the input window which are behind the output window.

An example of PKG index calculation is shown and described in Figure 1.

3 Problem formulation

To present our problem in a more formal way we prepared a model of a factory. In this model we treat the factory as a directed graph. Additionally, we consider tokens, which move through the graph. To make such a graph suitable for the factory we assigned separate meanings to nodes, arcs and tokens.

Nodes represent stationary objects, devices or places, which may delay the products and affect their sequence, like work stations, switches, buffers or quality control stations. Arcs represent flow of products between nodes in a fixed direction. This flow can be executed by transport lines or trucks between different plants. Finally, tokens represent product themselves. In a car factory which uses MTO system these products will be cars, car bodies or other parts which are produced or delivered in a sequence.

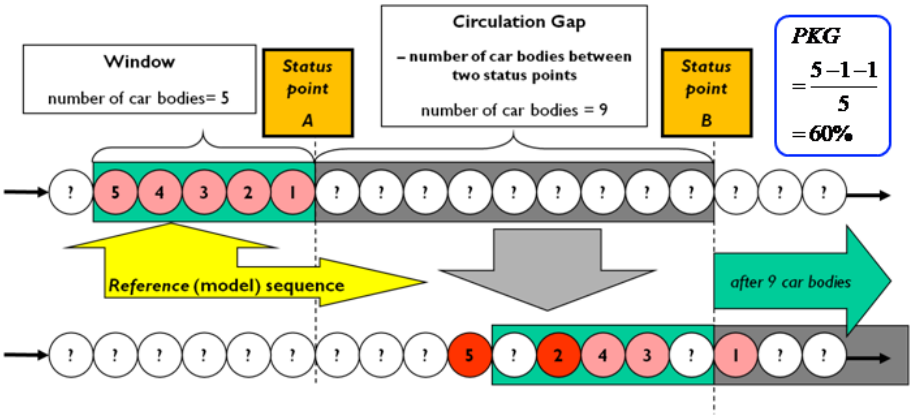


Figure 1: Example of PKG calculation. We use a window of size $P = 5$. Circulation gap G is equal to 9. In the output we can see one delayed car (number 2). Also one car is behind the window (5). Therefore, PKG index equals to 60% in this case.

Flow of products takes place according to some basic rules. Firstly, both nodes and arcs may contain a limited number of products at one moment. Sequence of products in every arc is preserved – products form a FIFO queue there. On the contrary, nodes may change the order of products they contain.

When we create a graph representing the factory, we assign specific functions to the nodes (e.g. planning nodes, buffer nodes) and arcs (e.g. transport lines). We can also distinguish the area of planning and the area of actual production. Tokens in the first one denote orders which are not any physical objects in the factory.

After creating a formal model of the factory we stay with a problem that most nodes require control. This especially concerns elements which may affect the sequence of cars in a significant way, like switches and buffers. Therefore, we want to answer the question: how can we create a system which controls a factory, modeled as a graph described above, according to some criteria?

In our solution we opt for a distributed control system. In this approach small devices control single nodes and arcs. Each device is responsible for a local area but it may communicate with other devices. The positive aspect of this solution is that most devices execute only simple algorithms. A distributed system is also easy to be adapted to new or irregular conditions. This allows an uncomplicated development of the factory. However, the approach has also some disadvantages. First of all, an infrastructure of many

small devices is required. But the main problem is that some devices have to make difficult decisions. This is the aspect which require thorough research. We began with creating a model which will help us program the devices so that they can make reasonable decisions.

4 Modeling

In our model we decided to use a multiagent system ([1, 2]) to control the factory, where each agent is responsible for a specific production area. One of the motivations for this approach is the possibility of integrating scheduling, control, monitoring and report systems by introducing several classes of agents. Moreover, rapidly changing demand for products in the production system requires a flexible solution for planning and control systems in the factory. The crucial aspects are also detection of different situations in the production environment and proper reaction of the control system. Time of this response determines the cost of preventing further damages in the system. It is specifically important in factories where control is dependent and complicated. An adaptive multiagent system can improve their general control and monitoring systems.

When developing a multilayer agent model, we execute the following plan. Firstly, we develop control scenarios and algorithms for low level production mechatronics hardware elements, like transportation lines, working stations, buffers, switches. These algorithms will be incorporated into local controllers like Programmable Logic Controllers (PLC), Industrial Computers (IC), Robot Controllers (RC). Then, we prepare communication and data exchange interfaces between low and high level control. Finally, we create interfaces between the production planning, management system and the floor shop control system.

To model procedures for the production system, we focus on particular production areas. The objective is to analyze the production process in the factory floor to find out the possibilities and necessities of the adaptive control and information system. In that process we can find specific places in the production system where the control or information system may build alternatives for the current way of controlling. When we find such "active points" in the particular production process, we may design types and basic functionalities of the agents.

For better communication agents are organized in a hierarchy which follows the hierarchy of factory [4]. It includes areas like plants, production lines and work stations. An example of such a hierarchy of agents is presented in Figure 2.

Type of an agent depends on a level in the hierarchy and a place in the factory where it appears. Some agents from the lower levels of the hierarchy

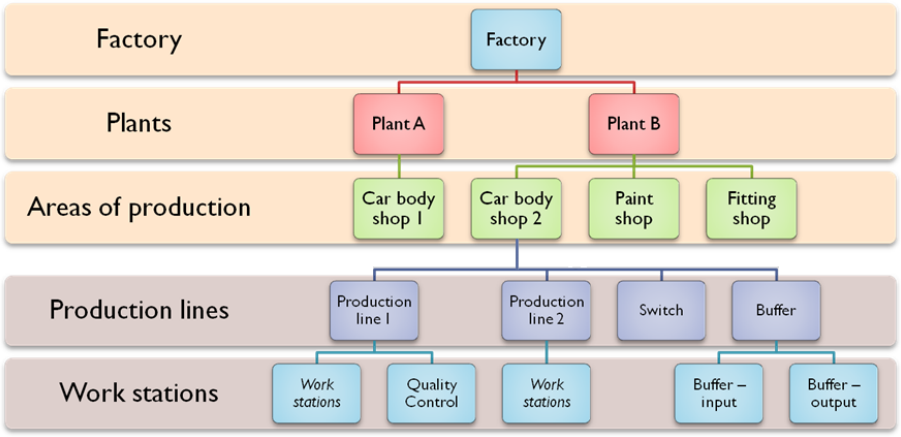


Figure 2: Exemplary hierarchy of agents.

refer directly to single nodes or arcs in a factory graph. Other agents refer to groups of agents which are below them in the hierarchy. An example of agent system which controls a specific factory is shown in Figure 3.

On every level of hierarchy we distinguish the following general types of agents:

- **Control agent** – decides the operation of the factory, directly or indirectly;
- **Monitoring agent** – collects information from a device or another agent and passes it to other agents;
- **Information agent** – calculates certain statistics and sends information outside the areas of production, e.g. to logistic system, supply chain, Just In Time (JIT) suppliers.

Implementation of an agent to the particular production system is characterized by parameters like domain of operation, input, output, work modes, state or AI Methods.

Algorithms for agents are designed according to some criteria, which they will optimize while the production. Possible general criteria for a factory are: indices of quality sequence (like PKG), stability of production, lateness/earliness, throughput rate, flow time. Current criterion optimized by an agent may be a part of the agent's work mode. We can also create a hierarchy of criteria by assigning ranks to them. For example, the main aim for an agent would be keeping the throughput rate on a reasonable level, but if there was enough time, the agent would also try to improve the quality of

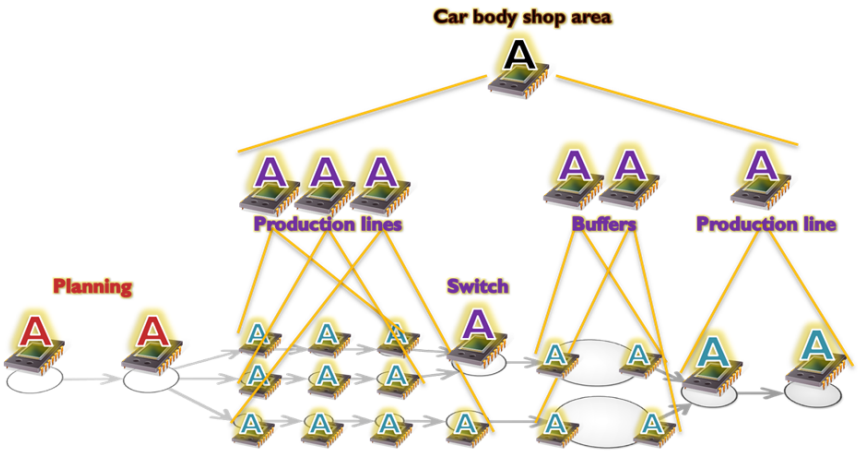


Figure 3: Fragment of an exemplary agent system controlling the factory.

sequence. Multiagent control system distributed on the lines, buffers and stations may automatically estimate the required production mode and prepare the system to the predefined modes.

5 Computational experiment

Testing the model presented in this paper would require simulating the whole factory and a multiagent system which controls it. This task is complex and remains as our goal. To test only a fragment of the agent system controlling one area of the factory we focused on the buffer management.

We considered a buffer which stores several parts of each car body, needed to build every car, and sends the sets of parts for each car directly to the merging station. Then, we implemented a buffer agent. It controls the most important action connected with buffer – selecting car bodies to be taken out of the buffers. We compared the operation of two algorithms for this action: a simple First In First Out (FIFO) method and the algorithm which aims at preserving the order from the status point where production actually begins. It turned out that the introduction of the second method, compared to the FIFO algorithm, caused a noticeable rise in PKG index. The important fact is that such a change affects only logic – no physical investments in the factory are needed.

6 Summary

Current centralized production control, which is often used in car factories, turns out to be inefficient. It cannot deal with anomalies and sequence preserving. We try to solve these problems and started with modeling the structure of a factory as a graph. Then, we proposed multiagent system for distributed control of the factory. Use of the agent system can adequately adapt to the fast changing production environment, including demand changes and production disturbances.

References

[1] Y. Shoham and K. Leyton-Brown. *Multiagent systems. Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

[2] J. M. Vidal. *Fundamentals of Multiagent Systems*. 2009.

[3] M. X. Weng, Z. Wu, and L. Zheng G. Qi. Multi-agent-based workload control for make-to-order manufacturing. *International Journal of Production Research*, 46(8):2197–2213, April 2008.

[4] J. Zhang, W. Zhai, and J. Yan. Multiagent-based modeling for reentrant manufacturing system. *International Journal of Production Research*, 45(13):3017–3036, July 2007.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

Simple assembly line balancing problem 2 with workers assignment - specification and solution

***Michał Morze*[1,2], *Grzegorz Pawlak*[2], *Alf Kimms*[3],**

***Tomasz Kujawa*[2], *Krzysztof Niespodziany*[2]**

1 Introduction

In this paper we introduce a version of a well-known Assembly Line Balancing Problem and propose three different solutions for this problem. We are considering an existing assembly line in the large car factory. The goal is to increase the production rate of the line (the number of assembled products per time unit) with respect to several constraints regarding some resource limitations.
The assembly line consists of a given number of stations where some operations (tasks) are performed. There is a set of workers, every of them is working on exactly one station. Every worker has qualifications (skills), which determines a set of operations he can perform. An operation is performed every *cycle time* by one worker on one station. The problem is to assign tasks to workers and allocate tasks to stations in order to maximize production rate with respect to the given precedence constraints (some tasks cannot be performed until some other are completed). By minimizing the cycle time we also minimize the sum of idle times for the stations.
This problem is a variation of Assembly Line Balancing Problem. Assuming the system produces only one type of a product at a time, the problem may be classified as Single-Model Assembly Line Balancing Problem [1]. We consider the problem deriving from the classical Simple Assembly Line Balanc-

[1]E-mail: michal.morze@student.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]University of Duisburg-Essen, Germany

ing Problem 2 (SALBP-2), which has the following characteristics:

- only one type (homogenous) product is being produced
- the production is paced (every *cycle time* one assembled product leaves the line)
- operation times are deterministic
- stations are homogenous
- the goal is to minimize the cycle time

2 Problem specification

The problem instance consists of the *m* stations, the precedence graph $\mathcal{G}$ and the set of *w* workers. Each worker has a set of operations he can perform (qualifications). The precedence graph is a non-cyclical digraph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V} = \{o_1, o_2, ..., o_n\}$ represents the set of operations (tasks) and $\mathcal{A} = \{(i, j) : i \in (V), j \in SUCCS(i)\}$ represents the precedence relation between operations. A set of immediate successors of the operation *i* is denoted by $SUCCS(i)$. The precedence graph provides a partial order of the operations. We minimize the cycle time *c* subject to the following constraints:

- Every tasks has to be assigned
- A task may be assigned to one worker on one station
- A worker may be assigned to at most one station
- The sum of times of operations assigned to a worker must not exceed the cycle time *c*
- For every operation *j*, where $(i, j) \in \mathcal{A}$, the operation *i* must be performed on earliest station than operation *j* or operations *i*, *j* must be performed by one worker on the same station

3 Solutions

We propose three different solutions for the SALBP-2 WA problem, including the linear programming model, the exact branch and bound algorithm and a heuristic algorithm.

3.1 Linear programming

Parameters

- n: The number of tasks.
- m: An upper bound for the number of required stations, e.g. $m = n$.
- o: The number of workers.
- E: Set of precedence constraints amnong the tasks.
- p_j: The processing time of task j.
- δ_{jh}: 1, if worker h is qualified to do task j. 0, otherwise.
- M: A big number, e.g. $M = n$.

Decision Variables

- τ: The cycle time.
- u_{jhk}: 1, if task j is performed by worker j at station k. 0, otherwise.
- z_{hk}: 1, if worker h has a positive workload at station k. 0, otherwise.

Programming Model Formulation

$$\min \tau$$

s.t.

$$\sum_{j=1}^{n} p_j \delta_{jh} u_{jhk} \leq \tau \qquad \begin{array}{l} h = 1, \ldots, o; \\ k = 1, \ldots, m \end{array}$$

$$\sum_{h=1}^{o} \sum_{k=1}^{m} k \delta_{ih} u_{ihk} \leq \sum_{h=1}^{o} \sum_{k=1}^{m} k \delta_{jh} u_{jhk} \qquad (i,j) \in E$$

$$\delta_{ih} u_{ihk} + (\sum_{h'=1}^{o} \sum_{k'=1}^{m} k' \delta_{ih'} u_{ih'k'} - \sum_{h'=1}^{o} \sum_{k'=1}^{m} k' \delta_{jh'} u_{jh'k'}) \leq \delta_{jh} u_{jhk} \qquad \begin{array}{l} (i,j) \in E; \\ h = 1, \ldots, o; \\ k = 1, \ldots, m \end{array}$$

$$\sum_{h=1}^{o} \sum_{k=1}^{m} \delta_{jh} u_{jhk} = 1 \qquad j = 1, \ldots, n$$

$$\sum_{j=1}^{n} \delta_{jh} u_{jhk} \leq M \cdot z_{hk} \qquad \begin{array}{l} h = 1, \ldots, o; \\ k = 1, \ldots, m \end{array}$$

$$\sum_{k=1}^{m} z_{hk} \leq 1 \qquad \begin{array}{l} h = 1, \ldots, o; \\ k = 1, \ldots, m \end{array}$$

$$u_{jhk} \leq \delta_{jh} \qquad \begin{array}{l} j = 1, \ldots, n; \\ h = 1, \ldots, o; \\ k = 1, \ldots, m \end{array}$$

$$u_{jhk} \in \{0,1\} \qquad \begin{array}{l} j = 1, \ldots, n; \\ h = 1, \ldots, o; \\ k = 1, \ldots, m \end{array}$$

$$z_{hk} \in \{0,1\} \qquad \begin{array}{l} h = 1, \ldots, o; \\ k = 1, \ldots, m \end{array}$$

$$\tau \geq 0$$

3.2 Branch and bound

We propose a branch and bound algorithm along with the following lower and upper bounds.

Definitions

Definition 1. *Earliest and latest stations*
Based on the maximal cycle time c earliest (Ej) and latest (Lj) station for each task

Zj are computed as follows:

$$E_j(c) = \left\lceil \frac{\sum_{k=1}^{PREDS(j)} max(PREDS_k) + t_j}{c} \right\rceil$$

$$L_j(c) = m + 1 - \left\lceil \frac{\sum_{k=SUCCS(j)}^{m} max(SUCCS_k) + t_j}{c} \right\rceil$$

Definition 2. *Station interval*
The station interval for each task is a range defined as:

$$SI_j(c) = \{E_j(c), E_j(c)+1, \ldots, L_j(c)+1, L_j(c)\}$$

Definition 3. *Temporary station interval*
The temporary station interval is a set of possible station assignments

$$tSI_j(c) = \{k : k \in SI_j(c) \wedge Z_j \in w(SI_j(c))\}$$

where $w(S_k)$ denotes the set of tasks possible to be done on the station k.

Definition 4. *sPPT (layers)*
Sets of potentially parallel tasks (layers) are defined as follows:

$$sPPT_i = \{z_j : z_j \textit{ has no predecessors in } sPPT_i\}$$

Each task belongs to exactly one sPPT.

Upper bound

$$UB_1 = max\{t_{max}, 2 \cdot \lfloor \tfrac{t_{sum}}{m} \rfloor\}$$

Proof of the correctness of this bound was presented by Hackman [2].

Lower bound

Lower bounds can be obtained from the formula

$$LB_1 = max\{t_{max}, \left\lceil \frac{\sum_{i=1}^{n} z_i}{\sum_{k=1}^{s} n(sPPT_k)} \right\rceil\}$$

where $n(sPPT_k)$ stands for the number of tasks in the layer k. This bound can be stated in terms of the related parallel machine scheduling problem. Processing of a job may be interrupted and continued on another machine - the task splitting is allowed.
A necessary condition for the existence of a feasible solution with m stations is $E'_j(c) \leq L'_j(c)$ for each task j. Otherwise, at least one task cannot be assigned to any station. Thus,

$$LB_2 = min\{c : L'_j(c) \geq E'_j(c) for j = 1, \ldots, n\}$$

defines a lower bound on the cycle time.
The critical path (a path in the precedence graph from any source to any sink containing the greatest sum of operation times) should be completed during the assembly process:

$$LB_3 = \left\lceil \frac{\sum_{z_j \in CP} z_j}{m} \right\rceil$$

Branch and bound algorithm

The following routine generate branches for the B&B algorithm.

1. Workers assignment – a partition of a set of *w* elements (workers) into *m* nonempty subsets (stations)
2. Calculating temporary station intervals (tSI)
3. Tasks to stations assignment — a single task is assigned to one station in each step
4. Tasks to workers assignment — a single task is assigned to one worker in each step

3.3 Heuristic algorithm

We propose the following heuristic approach for SALBP-2 WA problem:

1. Building sets of potentially parallel tasks (layers).
2. First estimation of workers to layers assignment. Greedy assignment provides first execution time estimation.
3. Deploying layers on stations — dynamic programming approach. Dynamic programming algorithm have been designed to deploy layers on stations minimizing cycle time difference between them in polynomial time. After this step feasible solution is constructed.
4. Optimization of workers assignment.

4 Conclusion

We defined the Simple Assembly Line Balancing Problem 2 with Workers Assigment and proposed three approaches to solve it. The future research will concern conducting a computational experiment to compare algorithms eficency and developing the solutions software visualisation tool.

References

[1] A. Scholl *Balancing and Sequencing of Assembly Lines.* Physica-Verlag, 1999.

[2] S.T. Hackman, M.J. Magazine, T.S. Wee Fast, Effective Algorithms for Simple Assembly Line Balancing Problems. *Operations Research*, 37:916–924, 1989.

Revenue maximization problem on parallel machines

***Jacek Juraszek*[1,2], *Erwin Pesh*[3], *Malgorzata Sterna*[2]**

1 Introduction

Revenue management is essentially the process of allocating resources to the right customer at the right time and the right price. A slightly different approach to revenue maximization can be met in "classical" scheduling theory, where the goal is to maximize the criterion value, i.e. the profit, for some given values of the problem parameters (cf. [6]). Such a model finds many practical applications. For example, a set of jobs can represent a set of customer orders which may give certain profit to a producer. Due to limited resources, modeled by a machine or a set of machines, the producer has to decide whether to accept or reject a particular order and how to schedule accepted orders in the system. Delays in the completions of orders cause penalties, which decrease the total revenue obtained from the realized orders. For this reason, maximizing revenue is strictly related to due date involving criteria such as minimizing tardiness or late work.

The maximum revenue objective function has been studied mostly for the single machine environment (cf. [2], [6], [7]). In our research, we investigate the problem of selecting and executing jobs on identical parallel machines in order to maximize the total revenue (profit) with the weighted tardiness penalty.

[1]E-mail: jacek.juraszek@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Information Systems, University of Siegen, Siegen, Germany

2 Problem Definition

We analyzed the scheduling problem of executing a set of jobs on a set of parallel identical machines. Each job is described by release time, which define moment of availability for processing by any machine. Processing time for each job is independent and know at the start of scheduling procedure. Moreover job processing can not be preempted. Any processed job have to end before deadline. There is also due date which define latest moment of finishing processing the job and getting for it full revenue. Exceeding due date cause tardiness of a job and revenue for processing such job is decreased by product of job weight and value of tardiness. Situation of decreasing revenue is modeling penalty or fine paid by the owner to client for meaningful but acceptable delay in ordered service. Every job can be assigned for processing to any machine, in other hand job can be rejected from schedule without any penalty. It is not "classical" ability of schedule problem, but it is modeling realistic situation when system owner have to choose most profitable set of orders (jobs).

NP-hardness of the problem of selecting jobs for execution with the tardiness penalty was proven by Slotnik and Morton [7] and also by Ghosh [2] who proved it for single machine problem. Problem intractability is considered since the problem of minimizing the weighted tardiness for a subset of accepted jobs, which is necessary to maximize the total revenue, is already hard [4, 5].

3 Solution Methods

In our research we proposed three strategy: branch and bound, list scheduling and simulated annealing. Only the first one is exact algorithm, remaining two are, accordingly, heuristic and metaheuristic.

A list scheduling algorithm was implemented in order to obtain an initial solution for simulated annealing as well as to determine an initial lower bound for the exact approach. Results also gives us point of reference in estimation of average solution quality of Simulated Algorithm, for instances of problem where optimal solution can not be reached because of unacceptable time consumption of branch and bound method. At each iteration, algorithm assigns an available job to a free machine, if such an assignment is feasible with regard to the job's deadline and if it will increase the overall revenue. Not assigned jobs are meant to be rejected. The method iterates over the set of jobs according to the priority dispatching rule.

The branch and bound algorithm constructs an optimal solution by analyzing all partitions of a set of jobs to two subsets of accepted and rejected ones. Then, it investigates all partitions of the accepted jobs to machines. For

a certain assignment of jobs, B&B checks all feasible permutations of jobs on machines. Constructing a partial schedule is suppressed, if the current revenue increased by the total revenue which might be obtained from non-assigned jobs (i.e. the upper bound of the criterion value) does not exceed the lower bound for the total revenue achieved so far.

We also propose a simulated annealing algorithm based on the classical framework of this method (cf. [3]) extended by local search procedure implemented as nested simulated annealing procedure. As schedule representation in SA algorithm we assume assignment each job to only one machine (including dummy machine containing set of rejected jobs). Transition from schedule representation to actual job schedule rely on exact approach. We developed two neighbor rules based on shifting or interchanging jobs between two assignment machines. New assignment was improved with local search. The cooling scheme is based on the geometrical temperature reduction. The initial temperature was determined by a tuning process. The method terminates after a given number of diversifications and iterations without of improvement.

4 Computational Experiments

Computational experiments were performed for 4 groups of randomly generated in- stances. There were instances with "tight" release times, in which most jobs were available for processing around time zero, and "loose" release times, which were spread in time. Similarly, instances with "tight" due dates and deadlines contained jobs with narrow time windows in which they can be processed without delay, while in instances with "loose" dues date and deadlines the lengths of these intervals were very close to the job processing times.

Test results disclosed the strong influence of the instance characteristic on the efficiency of the heuristic methods. Due to time requirements of the branch and bound algorithm, the comparison with optimal solutions was possible only for small instances. It showed the high efficiency of both heuristics, especially of simulated annealing, which generated mostly optimal solutions. simulated annealing found solutions with 98.73%, and list algorithm with 89.47% of the optimal revenue on average. In the computational experiments performed for large instances SA was still able to improve the quality of initial solutions generated by LA by 9.96% on average.

5 Conclusions

We investigated the problem of simultaneous selecting and scheduling a set of jobs on a set of identical parallel machines in order to maximize the total revenue and found it NP-hard. Due to problem intractability we designed metaheuristic algorithm and compared it to exact approach in the the extensive computational experiments. Within the future research, we will analyze a similar problem of orders acceptance and scheduling, which arises in a real world environment [1].

References

[1] L. Asbach, U. Dorndorf, E. Pesch Analysis, modeling and solution of the concrete delivery problem. *European Journal of Operational Research*, 193:820–835, 2009

[2] J.B. Ghosh Job selection in a heavily loaded shop. *Computers and Operations Research*, 24(2):141–145, 1997

[3] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983

[4] E.L. Lawler A pseudopolynomial algorithm for sequencing jobs to minimize total tardiness. *Annals of Discrete Mathematics*, 1:331–342, 1977

[5] J.K. Lenstra, A.G.H. Rinnooy Kan, P. Brucker Complexity of machine scheduling problems. *Annals of Discrete Mathematics*, 1:343–362, 1977

[6] H.F. Lewis, S.A. Slotnick Multi-period job selection: planning work loads to maximize profit. *Computers and Operations Research*, 29:1081–1098, 2002

[7] S.A. Slotnick, T.E. Morton Order acceptance with weighted tardiness. *Computers and Operations Research*, 34:3029–3042, 2007

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

Modelling SICMA problem as a problem of scheduling on FMS system

Michal Tanas[1,2,3,4]***, Witold Holubowicz***[2,3]***, Rafal Renk***[2,3]

Acknowledgments: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 217855 (SICMA Project).

1 Introduction

A scheduling problem is, in general, a problem answering a question of how to allocate some resources over time in order to perform a given set of tasks [1]. In practical applications resources are processors, money, manpower, tools, etc. Tasks can be described by a wide range of parameters, like ready times, due dates, relative urgency factors, precedence constraints and many more. Different criteria can be applied to measure the quality of a schedule.

Scheduling theory is widely applicable to solve real life discrete optimization problems. The purpose of this paper is to present another application of particular scheduling problems to the „Simulation of Crisis Management Activities" (in brief SICMA) EU project, whose objective is to improve health service crisis managers decision-making capabilities through an integrated suite of modelling and analysis tools providing insights into the collective behavior of the whole organization in response to crisis scenarios

[1]E-mail: michal.tanas@amu.edu.pl
[2]Applied Computer Science Division, Physics Faculty, Adam Mickiewicz University, PoznaĹ„, Poland
[3]ITTI sp. z.o.o., PoznaĹ„ Poland
[4]Institute of Computing Science, PoznaĹ„ University of Technology, PoznaĹ„, Poland

2 Description of the SICMA project

SICMA is EU FW7 Security Research Call 1 programme, whose point of interest is how to restore security and safety after a crisis. The term „crisis" denotes any unexpected and harmful event from the level of serious local accident (e.g. a car crash involving a truck carrying dangerous chemicals) to the level of state wide disaster (e.g. a large scale leak in a chemical factory, like Bhopal disaster in 1984). In order to enhance the efficiency of command and control by intelligent decision support systems, the task of the SICMA project is to develop appropriate novel approaches to computer assisted decision making. Applications should be robust and facilitate the cooperation of operational units across organizational boundaries. Two main research goals of the project are as follows:

1. The first challenge is to ensure that governments, first responders and societies are better prepared prior to unpredictable catastrophic incidents using new, innovative and affordable solutions.

2. The second challenge is to improve the tools, infrastructures, procedures and organizational frameworks to respond and recover more efficiently and effectively both during, and after, an incident.

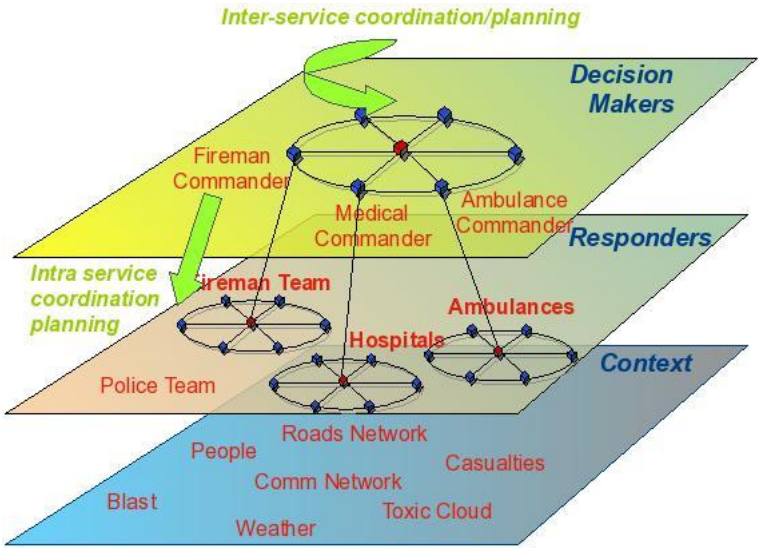
Decision-making support will be provided through an integrated suite of modelling and analysis tools. Key research aspects are:

- „bottom-up" modelling approach
- build independent model components and then combine them
- unpredictable factors modelling
- human behavior, mass behavior
- procedure support
- provide the user with the correct procedures to solve the problem
- computation of the "distribution" of the effectiveness of a certain "decision" rather than the effectiveness of that
- solution deterministically dependant on the preconceived scenario The combined effects of the above points will allow to document both the unexpected bad and good things in the organization(s) thus leading to better responses, fewer unintended consequences and greater consensus on important decisions.

3 Structure of crisis management services

In general the response to the crisis is the result of the activities of:

- different services (e.g. police, medical care, rescue forces, fire fighting, etc)
- interacting vertically (i.e. with components of the same organization) and horizontally (i.e. with components of other organizations)
- in a complex environment



4 Example crisis scenario

In context of the SICMA project many different crisis scenarios are considered. One of them is a scenario of industry accident or terrorist attack which results in a conventional explosion which releases a chemical agent and creates a large toxic cloud of chemicals inside a massively urbanized area. In this scenario the whole area of the crisis should be isolated, all unaffected people should be evacuated to create a safety buffer zone, and all victims (i.e. people injured by a explosion or toxic vapors) must be triaged and delivered to neighboring hospitals as fast as possible. The key problems in this scenario are:

- The largest hospitals are capable to take only 7 seriously injured patients per hour without risking of deteriorating quality of treatment. This ratio is surprisingly low.

- Victims must be delivered to hospitals by a transport system (i.e. ambulances) through very dynamic an unpredictable environment (i.e. unpredictable movement of chemical cloud, traffic jams, unpredictable crowd behavior, fires, building collapses, etc.)
- Each hospital has its own „menu" of treatments which it can perform. There are basic procedures which can be performed by any hospital, but there are some highly sophisticated medical procedures (i.e. decontamination of victims of not commonly used chemical agents) which can be performed only by a few hospitals in a state.
- First aid on site of disaster is performed by temporary hospitals called CCS (Casualty Clearing Stations) which similar to hospitals have very limited capacity.

5 Modelling SICMA as FMS

Considering all the facts above it is obvious that scheduling theory can be applied to reach the goals stated by EU. In particular in the example scenario presented above there are clear transformation from real-live crisis management problem into a scheduling problem, as follows:

- Hospitals and CCS'es can be modeled as subsystems containing several parallel processors.
- The number of such processors are equal to the number of casualties the hospital or CCS can simultaneously take care of.
- The „menu" of treatments a hospital can perform can be modeled by semispecialized processors, or by unrelated processors if infinite processing times are allowed.
- A casualty can be modeled by a task
- An evacuation of a hospitals (i.e. a hospital is on the way of the cloud) can be modeled by a processors breakdowns
- Triage and search and rescue teams activities can be modeled by a ready time
- Deteriorating of victims health in time can be modeled by due-dates
- Deaths caused by too late help can be also modelled by due-dates (this time with considerably higher penalty for late tasks)

- Ambulance service can be modeled by a transport system. Transport time is variable, there are various amount of different transport vehicles (e.g. ambulances, helicopters)
- Tasks are independent. Treatment of a casualty does not depend in any way on treatment of another casualty.
- Processing times are equal and tasks are identical. The project does not distinguish details like duration of particular medical procedures. Moreover, a casualty which was properly inserted into a hospital is no more object of interest of the SICMA (it is assumed that a hospital knows what to do). However the unability of a hospital to perform medical procedure suitable for a particual casualty may be modelled by the infinity processing time, so the processing times are either 1 or ∞

So, a real-life problem considered in the SICMA can be transformed a problem similar to two stage FMS problem, with ready times, transport, and with an unusual property that processors in machine center are unrelated not parallel and transport delays are variables

$$F2|p_{ij} = \{1, \infty\}, r_j, k_1, k_2, t_i = \{T_1 = f_1(t), \ldots, T_{k_2} = f_{k_2}(t)\}|U$$

Note, that in spite of the existence of *„exclusion lists"* which for each processor define set of tasks which cannot be processed on this particular processor, and so processors are not fully identical, the problem cannot be considered as job shop, because of the following key facts:

- Each job contains exactly one operation.
- There are several machines which can process a particular job.

6 Complexity of SICMA

The complexity of the SICMA problem is NP-hard, because SICMA is a generalization of $F2|p_{ij} = 1, t_i|C_{max}$ problem which was proven to be NP-hard by Wenci Yu in 1996 [2].

7 Relaxation of SICMA

Let us take some arbitrary assumptions which simpifies the original SICMA problem:

- Assumption that no traffic conditions are considered removes variability of transport times

$$F2|p_{ij} = \{1, \infty\}, r_j, k_1, k_2, t_i = \{T_1, \ldots, T_{k_2}\}|U$$

- Assumption that all hospitals and CCS are identical reduces the problem to a standard FMS problem with parallel processors in machine centers.

$$F2|p_{ij} = 1, r_j, k_1, k_2, t_i = \{T_1, \ldots, T_{k_2}\}|U$$

- Assumption that all casualties are already found and triages removes ready times from the problem.

$$F2|p_{ij} = 1, k_1, k_2, t_i = \{T_1, \ldots, T_{k_2}\}|U$$

Let us denote such simplified SICMA problem, where there are k_1 CCS and k_2 hospitals as $SICMA - k_1, k_2$. The complexity of $SICMA - k_1, k_2$ problem is still open.

8 Conclusion

Among many others practical applications, scheduling theory may be successfully used in health care decision support systems and in modeling crisis management activities, ambulance service in real urbanized areas environment. More over the SICMA EUs project creates several new detailed problems in domain of scheduling theory, which needs to be solved.

References

[1] K. Baker. *Introduction to Sequencing and Scheduling*. J. Wiley, New York, 1974.

[2] W. Yu. *The Two-machine Flow Shop Problem with Delays and the One-machine Total Tardiness Problem*. Technische Universiteit Eindhoven, 1996.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

Variable sized bin packing problem

***Piotr Gawron*[1,2], *Rafal Walkowiak*[2]**

1 Introduction

The issue of Variable Sized Bin Packing Problem is very common in different fields of research. It can be transformed to many other well-known problems, for example:

- truck loading problem (selection of trucks for a given load)
- parallel processing scheduling (cost optimization by selecting parallel machines capable to solve the set of problems)

The problem belongs to NP-hard class, so there are many heuristic approaches proposed for solving it (see [2]). The authors proposed an exact method, which can be applied for small instances. Moreover the way of improvement for bigger instances was shown.

2 Problem Formulation

The issue is an extended version of bin packing problem which is widely considered by many authors (see [3]). Briefly, it can be defined as follows:

- there is a set of items, each one with a certein weight
- there is a set of bins, each one with a given positive capacity and a cost
- the goal is to pack all the items in the bins and minimizing the total cost associated with the chosen bins

The formal definition of the problem under discussion is an Integer Programming definition:

[1]E-mail: piotr.gawron@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland

- data:

  $n -$ number of items, maximum number of containers
  $s_j -$ size of item j
  $K -$ number of types of containers
  $L_k -$ max. size of container which has lower cost than container k
  $U_k -$ size of container k
  $c_k -$ cost of container k filled with items

- decision variables:

$$x_{ij} - \begin{cases} 1 \text{ if the item j is assigned to container i} \\ 0 \text{ otherwise} \end{cases}$$
$$y_{ik} - \begin{cases} 1 \text{ if the container i is of type k} \\ 0 \text{ otherwise} \end{cases}$$

- objective function:

$$min \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} c_k$$

- constraints:

$\sum_{i=1}^{n} x_{ij} = 1$	$j = 1 \dots n$	every item j is assigned to a single container
$\sum_{j=1}^{n} x_{ij} c_j \leq \sum_{k=1}^{K} y_{ik} U_k$	$i = 1 \dots n$	every container i filled with items gets suitable size
$\sum_{k=1}^{K} y_{ik} \leq 1$	$i = 1 \dots n$	at most one size is assigned to every container i

$x_{ij}, y_{ij} \in \{0, 1\}$

3 Method

To solve this problem we propose an exact solution which is based on dynamic programming. In our model we memorize the best solution for every subset of input items. In Table 3 the output of the computation is shown for very small example input (4 items). As we can see, the memory complexity of our algorithm is exponential: $O(n^2)$. This means that for 30 elements, we need gigabytes of memory, but for forty elements, the memory is counted in terabytes. However, its worth to notice that not all of the memory cells are used to compute the final solution. In considered example cell containing information about subset consisted only from item 1 is worthless, because used bin have enough place to put there item 4 and this place is wasted in such situation. In fact around 99% of cells contains such unimportant information. Another possibility to reduce memory usage is to divide problem

into two subproblems and as a solution give assembly of solutions to these subproblems. Time complexity of our algorithm is $O(n^4)$. However it is hard to find a case which compute so slow.

To solve the above problem, an exact algorithm based on dynamic programming was proposed. In the algorithm the best solution for every subset of input items was stored. The output of the computation for very small input was shown in Table 3. The memory complexity of the algorithm is exponential: $O(2^n)$. This means that for 30 elements, gigabytes of memory is needed, but for 40 elements the memory is counted in terabytes. However, it is important to stress the fact that not all of the memory cells are used in computing the final solution. In the considered example, cell containing information about subset consisting only of item 1 is worthless. The used bin has enough space to put there item 4 and this space is wasted in this situation. In fact, around 99% of cells contain such unimportant information. Another possibility to reduce memory usage is to divide the problem into two subproblems. As a result, assembly of the solutions was taken into account. Time complexity of the algorithm is $O(4^n)$. However, it is hard to find a case which is computed by the algorithm so slowly.

The proposed solution of variable sized bin packing problem is still being developed. The main obstacle is memory limit. The authors are currently trying to reduce the execution time by cuting off some useless computations.

Item number	Size
1	1
2	2
3	5
4	6

Table 1: Input items of the example.

Bin size	Cost
3	4,5
7	7,0

Table 2: Types of bins of the example.

References

[1] I. Correiaa, L. Gouveiab, F. Saldanha-da-Gamab *Solving the varaible size bin packing problem with discretized formulations.* 2008.

Item 1	Item 2	Item 3	Item 4	Min cost	Assignment
0	0	0	0	0	-
1	0	0	0	7	(1)
1	0	0	1	7	(1,6)
1	0	1	0	7	(1,5)
1	0	1	1	11,5	(1,5)(6)
1	1	0	0	4,5	(1,2)
1	1	0	1	11,5	(1,2)(6)
1	1	1	0	11,5	(1,2)(5)
1	1	1	1	14	(1,6)(2,5)

Table 3: Output of computation from example input (1 in first 4 columns means that appropriate item is used).

[2] M. Haouari, M. Serairi *Heuristics for the variable sized bin-packing problem.* 2009.

[3] C. C. Lee, D. T. Lee *A simple on-line bin-packing algorithm.* 1985.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

A computational tool for identifying putative cis-regulatory modules

Klaus Ecker[1,2]

1 Introduction

It is widely accepted knowledge that the expression of a gene is controlled by "regulatory switches" that are realized by certain proteins called transcription factors which are able to bind on short length DNA sections. Binding sites for gene expression are mostly located upstream in non-coding genomic areas, often close to the transcription start site. Identifying binding sites is an important step in the study of gene regulation mechanisms. As experiments are time consuming and expensive, applying computational search methods was appealing and led to the development of over 100 computational tools for discovering binding sites. Among the methods realized in these tools, essentially two different principles can be distinguished. The larger group applies stochastic analysis methods to reveal words (motifs) of putative biological interest. These methods are based on the hypothesis that words having unusual structure or being over- or underrepresented from a statistical point of view should have biological meaning. Typical methods search maximally expected words (as has been done, for example, in MEME [1, 2]), words with maximum likelihood ([3]), or implement a Gibbs sampling strategy ([4, 5]). The other class of motif discovery intends to identify motifs by comparing sequences that were expected to share common motifs. These methods perform a search for common similar words in two or more sequences, where similarity of words is defined by means of Hamming or edit distance. Examples of tools using Hamming distance are YMF [6], WEEDER [7], and SiteSeeker [8].

It is well known (viz. e.g. [9]) that motifs are usually members of a com-

[1]E-mail: ecker@ohio.edu
[2]Center for Intelligent, Distributed and Dependable Systems, Ohio University, Athens, Ohio, USA

plex network for regulating the expression of genes. Over a cascade of several steps, regulatory genes synthesize transcription factors that control the expression intensity of other regulatory elements or target genes.

One important objective in the attempt to understand mechanisms of gene regulation is the elucidation of the network structure. A cis-regulatory module (CRM) is understood as part of the genome that comprises a set of short length binding sites of a regulatory network. Identifying the transcription factor binding sites of a network, i.e., analyzing the structure of a module will obviously be the preliminary step of a detailed network analysis [10]. Experimental investigation revealed that functional regulatory elements are often organized in relatively dense clusters, covering a stretch of a few hundred bases [11, 12, 13, 14, 15, 16, 17]. These cis-regulatory modules are usually found closely located to the basal transcription apparatus of the target gene, but may also be as far as several kilo-bases away [18].

As compared to motif discovery, the problem of computationally identifying CRMs is much more difficult – of course mainly because they are combinations of binding sites. Further difficulties arise from the observation that CRM components appear to be rather short (≥ 6 base pairs) and often degenerated [19].

On the other hand, gene regulatory systems turn out to be quite stable during evolution, as compared to relatively frequent replication processes of genes and mutations of the coding sequences [9, 12]. This conservation property of regulatory code can advantageously be used by computational methods for identifying cis-regulatory modules of potentially co-regulated genes.

Two fundamentally different approaches can be distinguished:

(a) Searching single sequences for clusters of motifs. Probabilistic criteria for motif selection, such as the log likelihood ratio or the information content [3, 20, 21, 22, 23] allow discriminating motifs from the background distribution. Early solutions for the CRM discovery problem followed the concept that motifs in a module may already be known. Accordingly these methods start from known motifs (MSCAN [14], ModuleSearcher [18], ClusterBuster [19], Cister [20], CREME [26], Stubb [27], Composite Module Analyst [28]), or from putative motifs found with the aid of motif discovery tools (PFRSearcher [12], Gibbs Module Finder [16], Cis-Module [29], Target Explorer [30], EMCModule [31], CSam [32]). Significant improved prediction success was gained by a holistic view that centered on clusters of binding sites instead of isolated motifs, see [32] for an overview. Following this approach, a cis-regulatory module is modeled as a window in a non-coding DNA region which has a high density of binding sites. In addition, the structure of and the relationship between the binding sites of a module should be captured as precisely as possible. Structural conditions may regard, for example, the stretch of a cluster and the number of binding sites [13, 16, 27, 32, 33], the

lengths and overrepresentation of binding sites [17, 32, 34, 35], the distances between sites, and the density of binding sites.

(b) Identifying clusters by comparing two or more promoter sequences from gene batteries or from phylogenetically related genomes. This approach is guided by the hypothesis that genes being expressed under similar conditions should be controlled by similar regulatory networks, and hence should share similar cis-regulatory modules. This leads to the question of how to measure the similarity of modules. Sequence alignment methods like BLAST detect similar structures in biological sequences. However, sequences even with common modules can still have a rather weak similarity, as the matching sites are short, may be degenerated, and do not necessarily appear in the same order. An accurate definition should capture information about the binding sites themselves, such as their similarity, the distances between matching pairs of sites, their order in the modules, etc.

2 The CRMODULE tool

Following the comparison paradigm *(b)*, we developed a tool for computationally identifying common modules in pairs of sequences. The method uses a number of parameters, (*i*) for defining modules, and (*ii*) for comparing two modules. A module is generally defined as a cluster of binding sites. Parameter class (*i*) defines restrictions such as upper bounds for the span (or module stretch), minimum and maximum distances between binding sites (gap lengths), maximum allowed overlaps of binding sites, and density of binding sites in the cluster. Parameters (*ii*) restrict the differences of parameters values, such as span difference, the gap variation between matching pairs of sites, and the degree of order conservation.

This concept has been implemented in the tool CRMODULES[1] [36]. It is motivated by the idea that the user should be allowed to define restrictions for the similarity attributes in accordance with practical experience. In other words, the definition of "module similarity" lies completely in the hands of the user. CRMODULES hence returns only those common modules that follow the given user-defined restrictions.

CRMODULES implements the following algorithm.

Input: Two DNA sequences S_0 and S_1 that are suspected
to have common regulatory regions.
Set of restriction parameters:
minimum length of binding sites (l_{min}),
maximum allowed Hamming distance between matching words,

[1] CRMODULES can be accessed at `http://siteseeker.msseeker.org/`

the module definition parameters
maximum module span ($span_{max}$)
lower and upper bounds for the gaps between neighboring sites
minimum density of sites,
and the module similarity parameters
maximum difference of modules spans,
variation of gap lengths of pairs of matching words.
percentage of order conservation of matching words,

Output: Set of modules satisfying the module conditions

The algorithm is organized as follows (viz. Fig. 4).

Phase 1: In S_0 and S_1, find all pairs of common binding sites of lengths $\geq l_{min}$.

Phase 2: A sliding window of length $span_{max}$ scans the list of matching pairs, and prints the modules satisfying the given module conditions.

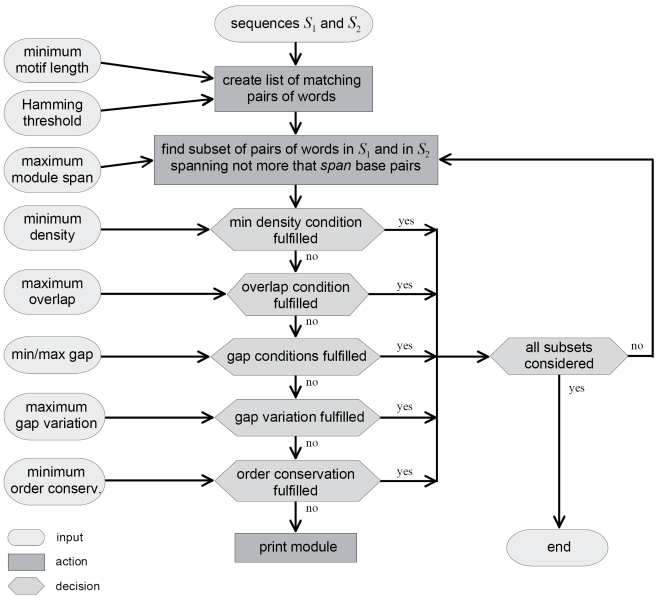


Figure 4: Flow diagram of CRMODULES

Time complexity analysis: Phase 1 finds all similar pairs (w_1, w_2) of length

l_{min} , $w_1 \in S_1$, $w_2 \in S_2$. The needed time is $O(|S_1| \cdot |S_2|)$, where $|S_i|$ denotes the length of sequence S_i . Suppose the list has k word pairs, where obviously k is upper bounded by $|S_1| \cdot |S_2|$. By sliding through the list of word pairs, phase 2 finds all clusters of words that fit in the frame of size $span_{max}$ which takes $O(k)$ time. Each cluster is checked against the modules conditions. For each cluster, the algorithm has to go through all the word pairs whose number n is upper bounded by $span_{max}$. Only checking the order condition is of exponential time complexity $O(2^n)$. All other conditions take $O(n)$ time. So, in total, the algorithm takes $O(|S_1| \cdot |S_2| + n2^n)$ steps, where n is an upper bound for the number of words in a module.

In practice it turns out that the exponential part consumes insignificant time because modules are usually very small. The experience with CRMODULES shows that, for sequences of lengths 1000, maximum span = 60 and word length 6, the first phase takes approx. 10 times longer than the second phase.

3 Experimental studies

Of particular interest is searching for identical modules with non-overlapping binding sites in two sequences. We performed extensive investigation on genes encoding proteins of the *Arabidopsis thaliana* f-box family. The f-box protein family is known to play a crucial role in protein-protein interaction [37]. Because of space limitation we show the outcome for only one pair, i.e., the sequences At2g07140 and At3g44120 with respective promoter lengths 1157 and 1417. The search conditions were as follows: "Find all modules with at least four exactly matching binding sites, each of length at least 8 base pairs, and the same gap lengths between each pair of matching sites." The output shows the three modules (in fact they can be combined to a single module of 6 sites):

```
At2g07140.1:  TAAGATGTTGTATTGTCAGATGTTTCAA ... TGCTCAGTATT ... TTCCTTGC ... TTAAACACA
  Positions:  -613                            -584           -572         -563
At3g44120.1:  TAAGATGTTGTATTGTCAGATGTTTCAA ... TGCTCAGTATT ... TTCCTTGC ... TTAAACACA
  Positions:  -1161                           -1132          -1120        -1111
```

```
At2g07140.1:  TGCTCAGTATT ... TTCCTTGC ... TTAAACACA ... AAAGTTTATGATCAAA
  Positions:  -584            -572         -563          -547
At3g44120.1:  TGCTCAGTATT ... TTCCTTGC ... TTAAACACA ... AAAGTTTATGATCAAA
  Positions:  -1132           -1120        -1111         -1095
```

```
At2g07140.1:  TTCCTTGC ... TTAAACACA ... AAAGTTTATGATCAAA ... CTCTTCTTGTAGAG
  Positions:  -572         -563          -547                 -530
At3g44120.1:  TTCCTTGC ... TTAAACACA ... AAAGTTTATGATCAAA ... CTCTTCTTGTAGAG
  Positions:  -1120        -1111         -1095                -1078
```

The result is quite unusual because of the length of words and the identical gap lengths for each matching pair of sites. From this one may conclude that the found module plays indeed an important role in the regulation of the two genes. The unusual character was further verified by a random analysis where, following the same module restrictions, 1000 samples (each sample is a pair of randomly generated sequences) were analyzed for common modules. It turned out that probability of common modules with two sites is only 0.0030; no common module with 3 or more sites were found.

4 Summary

We motivated and described a computational method for the *de novo* discovery of modules common in two DNA sequences. This method reveals modules solely by a comparison paradigm, and performs the search independently of information about known motifs. A number of simple parameters are available for which tolerance bounds can be specified. This offers the user a high degree of flexibility to define what should be understood by module similarity in a particular practical situation. A prototype tool CrModules implementing the proposed features is described and an application on real data discussed.

References

[1] Bailey T. L., Williams, N., Misleh, CH. and Li, W. W. (2006), MEME: discovering and analyzing DNA and protein sequence motifs, *NAR 34*, 2006, W369-W373.

[2] Bailey T. and Elkan C. (1994), Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.2*, 28-36.

[3] Hertz, G. and Stormo, G. (1999), Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, *Bioinformatics* **5**, 563-577.

[4] Thompson W. A., Newberg L. A., Conlan S., McCue L. A. and Lawrence C. E. (2007), The Gibbs Centroid Sampler, *Nucleic Acids Res.*, **35**(Web Server issue): W232–W237.

[5] Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C., (1993), Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science* **262**(5131), 208-214.

[6] S. Sinha , M. Tompa, Discovery of novel transcription factor binding sites by statistical overrepresentation, *Nucleic Acids Res.* 30, 5549-5560, 2002.

[7] Pavesi G., Mereghetti P., Mauri G., Pesole G. (2004), Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes, *Nucleic Acids Res*. 32 (Web Server Issue), W199-203.

[8] Ecker, K., Lichtenberg, J. and Welch, L. (2008) The SiteSeeker Motif Discovery Tool, *In Silico Biology* 9, 0002.

[9] Davidson, E. H. (2006). The Regulatory Genome. Gene Regulatory Networks in Development and Evolution. Elsevier.

[10] Klepper, K., Sandve, G. K., Abul, O., Johansen, J. and Drablos, F. (2008). Assessment of composite discovery methods, *BMC Bioinformatics* 9, 123.

[11] Davuluri, R. V., Sun, H., Palaniswamy, S. K., Matthews, N., Molina, C., Kurtz, M. and Grotewold, E. (2003). AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors, *BMC Bioinformatics* 4 ,25.

[12] Grad, Y. H., Roth, F. P., Halfon, M. S. and Church, G. M. (2004). Prediction of similarly acting cis-regulatory modules by subsequence

profiling and comparative genomics in Drosophila melanogaster and D.pseudoobsuca, *Bioinformatics* 20, 2738-2750.

[13] Davidson, E. H. (2001). Genomic Regulatory Systems: Development and Evolution. Academic Press, San Diego.

[14] Johansson, Ö., Alkema, W., Wasserman, W. W. and Lagergren, J. (2003). Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm, *Bioinformatics* 19 Suppl 1, 169-176.

[15] GuhaThakurta D. and Stormo, G. D. (2001). Identifying target sites for cooperatively binding factors, *Bioinformatics* 17, 608-621.

[16] Thompson, W., Palumbo, M. J., Wasserman, W. W., Liu, J. S. and Lawrence, Ch. E. (2004). Decoding Human Regulatory Circuits, *Genome Res.* 14, 1967-1974.

[17] Kantorovitz, M. R., Robinson, G. E. and Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences, *Bioinformatics* 23 (ISMB/ECCB), i249-i255.

[18] Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. and De Moor, B. (2003). Computational detection of cis-regulatory modules, *Bioinformatics* 19 Suppl 2, 5-14.

[19] Frith, M. C., Li, M. C. and Weng, Z. (2003). Cluster-Buster: finding dense clusters of motifs in DNA sequences, *Nucleic Acids Res*. 31, 3666-3668.

[20] Frith, M. C., Hansen, U. and Weng, Z. (2001). Detection of cis-element clusters in higher eukaryotic DNA, *Bioinformatics* 17, 878-889.

[21] Frith, M. C., Spouge, J. L., Hansen, U. and Weng, Z. (2002). Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*. 30, 3214-3224.

[22] Crowley, E. M., Roeder, K. and Bina, M. (1997). A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol*. 268, 8-14.

[23] Rajewsky, N., Vergassola, M., Gaul, U. and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* 3, 30.

[24] Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl Acad. Sci. USA* 99, 757-762.

[25] Rebeiz, M., Reeves, N. L. and Posakony, J. W. (2002). SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl Acad. Sci*. USA 99, 9888-9893.

[26] Sharan, R., Ovcharenko, I., Ben-Hur, A. and Karp R. M. (2003). CREME: a framework for identifying cis-regulatory modules in human–mouse conserved segments. *Bioinformatics* 19 (Suppl. 1), i283-i291.

[27] Sinha, S., van Nimwegen, E. and Siggia, E. D. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics* 19 (Suppl. 1), i292-i301.

[28] Kel, A., Konovalova, T., Waleev, T., Cheremushkin, E., Kel-Margoulis, O. and Wingender, E. (2006). Composite module analyst: a fitness-based tool for identification of transcription factor binding site combinations, *Bioinformatics* 22, 1190-1197.

[29] Zhou, Q. and Wong, W. H. (2004). CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling, *PNAS* 101, 12114-12119.

[30] Sosinsky, A., Bonin, C. P., Mann, R. S. and Honig, B. (2003). Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors. *Nucleic Acids Res*. 31, 3589-3592.

[31] Gupta, M. and Liu, J. S. (2005). De novo cis-regulatory module elicitation for eukaryotic genomes, *PNAS* 102, 7079-7084.

[32] Ivan, A., Halfon, M. S. and Sinha, A. (2008). Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs, *Genome Biology* 9, R22.

[33] Carroll, S. B., Grenier, J. K. and Weatherbee, S. D. (2004). From DNA to Diversity, Molecular Genetics and the Evolution of Animal Design, Blackwell Publishers, Oxford.

[34] Pierstorff, N., Bergman, C. M. and Wiehe, T. (2006). Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics* 22, 2858-2864.

[35] Li, L., Zhu, Q., He, X., Sinha, S. and Halfon, M. S. (2007). Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses, *Genome Biology* 8, R101.

[36] Ecker, K. and Welch L. (2009). A concept for ab initio prediction of cis-regulatory modules, accepted for publication in *In Silico Biology*.

[37] Kipreos, E. T. and Pagano M. (2000), The F-box protein family, *Genome Biology* 2000; 1(5): reviews3002.1–reviews3002.7.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

Protein structure modelling – case study

Maciej Milostan [1,2]***, Joanna Sarzynska*** [3]***, Agnieszka Mickiewicz*** [3]***, Maciej Antczak*** [2]***, Piotr Lukasiak*** [2,3]***, Jacek Blazewicz*** [2,3]

1 Introduction

Proteins are one of the most important building blocks of life, they catalyze the necessary reactions in the cells to sustain life and improve metabolism. They are very complicated molecules, which in many cases fulfill exquisitely specific task. We are witnesses of fast improvement of the techniques for identification of protein sequences. Thus the number of such sequences gathered in databases (e.g. PDB) increased tremendously but only for the fraction of them the three dimensional structure is known.

In our current research we are interested in solving the structure of dicer like proteins in Arabidopsis Thaliana in order to better understand processes observed during wet-lab experiments. Dicer like proteins are very important for the process of micro-RNA creation. At the moment there is only one structure of dicer, from Giardia Intestinalis, determined experimentally via crystallography [1]. This fact makes the secrets of dicer like proteins harder to reveal by means of comparative modelling. However we managed to provide several interesting models and we would like to share here our experience.

[1]E-mail: maciej.milostan@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

2 Problem Formulation

Dicer in Giardia Intestinalis [1] cuts 25bp long fragments of RNA. In Arabidopsis Thaliana there are four dicer like proteins (DCL) that cuts 21 and 24bp long structures. The main goal is to determinate differences between them based on models, however here we focus only on selected aspects of protein structure models construction.

The input for the problem is structure of known dicer, sequences of dicer like proteins from Arabidopsis Thaliana and other organisms, databases of domains, set of templates for each well known domain. The structure of dicer is shown in Figure 5.

The output is structure of each DCL like protein from Arabidopsis. An example of such a structure is shown on Figure 6.

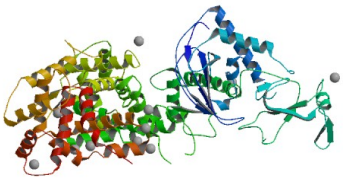


Figure 5: Structure of Dicer from Giardia Intestinalis.

3 Method

Methodology that we used is based on homology modelling paradigm. Thus we can distinguish following steps in our procedure:

- Identification of conserved domains using CDD-Search [7].
- Identification of template structures for identified domains by psi-blast [3] and literature studies.

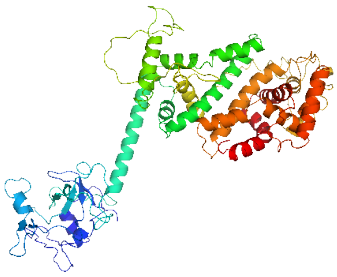


Figure 6: Structure of Dicer Like (DCL) Protein from Arabidopsis.

- Preparation of multiple sequence alignments for all DCLs, for each domain separately, and their template structures using ClustalW [4]. Additionally, all templates for each domain has been aligned using stamp algorithm from VMD tool.
- Manual concatenation of domain based alignments.
- Adjustment of the multiple sequence alignments using expert knowledge about conserved regions and binding sites.
- Detection of missing sequence fragments in crystallized structure of the template and incorporation of this information into alignments.
- Improvement of the alignment using secondary structures prediction obtained from JPred3 server [6].
- Generation of tertiary structures using Modeller [2] from Sali's Lab[1].
- Docking RNA using Haddock [5].

4 Results

Based on protocol mentioned above we have generated several models of the Dicer Like protein fragment and run preliminary docking experiments. As

[1] `http://www.salilab.org/modeller/about_modeller.html`

a result we have obtained interesting structure of protein-RNA complex. At the moment we are investigating properties of our models trying to understand were lies the crucial differences in DCLs. One of our results has been illustrated in the Figure 7.

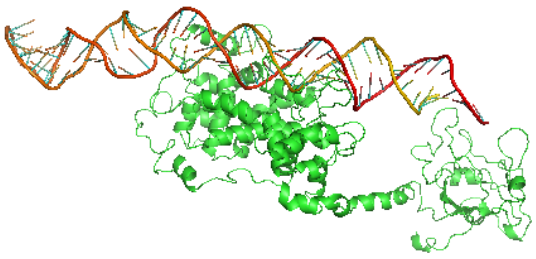


Figure 7: Preliminary result of RNA docking to modelled structure of dicer like protein

5 Conclusion

In conclusion, we can speak that our methodology of modelling proven to be correct and we have accomplished a first basic step toward better understanding of the way how dicer like proteins function in plants. The work is still in progress and we believe that further work will be successful

References

[1] I.J. MacRae, K. Zhou, F. Li, A. Repic, A.N. Brooks, W.Z. Cande, P.D. Adams, J.A. Doudna, Structural Basis of Double-Stranded RNA Processing by Dicer *Science*, 311:195-198, 2006.

[2] M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29:291-325, 2000.

[3] S.F. Altschul, T.L. Madden, A.A. Schäffer1, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman Gapped BLAST and PSI-BLAST: a new generation of protein database search programs *Nucleic Acids Research*, 25(17):3389-3402, 1997.

[4] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947-8, 2007.

[5] S.J. de Vries, A.D.J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar and A.M.J.J. Bonvin HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins: Struc. Funct. & Bioinformatic*, 69:726-733, 2007.

[6] C. Cole, J.D. Barber and G.J. Barton The Jpred 3 secondary structure prediction server *Nucleic Acids Research*, 36(Web Server issue):W197–W201, 2008.

[7] A. Marchler-Bauer , J.B. Anderson, F. Chitsaz, M.K. Derbyshire, C. DeWeese-Scott, J.H. Fong, L.Y. Geer, N.R. Gonzales, M. Gwadz, S. He , D.I. Hurwitz, J.D. Jackson, Z. Ke, G.J. Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, GH Marchler, M. Mullokandov, J.S. Song, A. Tasneem, N. Thanki, R.A. Yamashita, D. Zhang, N. Zhang, S.H. Bryant. CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, 37:D205-D210, 2009.

Graphs in NMR analysis of RNAs

Marta Szachniuk[1,2,3]*, Mariusz Popenda*[2]*, Lukasz Popenda*[4]

Acknowledgements: The work has been partially supported by NN519314635 grant from the Ministry of Science and Higher Education, Poland.

1 Introduction

Graphs are mathematical structures widely used to represent relations between different objects. In computer science they are often found as datum points in the design of algorithms solving problems of various types. Thus, defining a good graph model is often a very important aspect of dealing with a novel question. It allows for a theoretical analysis of the problem, including its computational complexity, differentiation between the ideal and real version, defining exceptions, introducing new methods to solve the problem. Many bioinformatic applications rely on graph models. Especially structural bioinformatics provides a vast range of issues which can be modelled via graph theory. Studying the problems of structure determination, prediction or comparison has already resulted in a creation of new graph models to represent experimental data and structures, e.g. protein graphs [7], DNA graphs [2], RNA graphs [4], NMR graphs [5], etc. These models have a great influence on further computational processes. It is also clear that

[1]E-mail: marta.szachniuk@cs.put.poznan.pl
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
[3]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[4]Department of Systems and Computer Science, University of Florence, Florence, Italy

topology of the graphs depends on the features of the associated data, like molecule shape or interacions found between the atoms.

In this paper we discuss the ideas concerning a process of a determination of RNA structures on the basis of NMR spectroscopy. Nuclear Magnetic Resonance gives a unique opportunity to study molecules in solution which is similar to their natural environment. The experiments provide the information about the molecule in action and result in obtaining a family of conformers. The determination of these structures is a multi-stage process starting from an experimental part, in which multidimensional correlation spectra are acquired. The quality and quantity of these spectra strongly influence the next, computational part, where the following steps are accomplished: processing, peak-picking, assignment, restraints determination, structure generation and refinement [1]. The procedure of assigning the observed signals to the corresponding protons and other nuclei is a bottleneck of the structure elucidation process. For non-labeled small proteins, as well as short DNA and RNA duplexes, the assignment of NMR signals is usually based on the analysis of two dimensional spectra like NOESY, TOCSY and COSY. For more complex structures it is necessary to use 3D and 4D spectra. Here, we model the sequential assignment problem in 2- and 3-dimensional space of correlation signals observed for RNA molecules.

2 Method

The results of NMR experiment performed for the molecule are given in the spectral form. Each interaction occurring between two atoms during the experiment is represented as a cross-peak in the NMR spectrum. The cross-peak is characterized by the coordinates of its centre, widths in all the dimensions and a volume of the interaction. Each coordinate of the cross-peak centre defines a position (chemical shift) of one atom participating in the corresponding interaction. The above mentioned features of the cross-peaks and thus, of the signals are known from the experiment. However, no information is provided about the atoms which generated each signal. Thus, a necessary first step is to map the spectral data to atoms in the molecule sequence and determine the resonance assignment. In case of an analysis of

short RNA sequences, assignment procedure starts from a reconstruction of a transfer pathway, called NOE pathway, in the 2D NOESY spectrum. The NOE pathway is known to connect cross-peaks for H8-H1' and H6-H1' interactions, which are located in the aromatic/anomeric region of the spectrum. The pathway crosses peaks for intra- and internucleotide signals alternately and its length is obviously related to the number of H1', H6 and H8 atoms in the molecule structure. In [1] we have proposed the first theoretical graph model of the assignment problem in the NOESY spectrum recorded for RNA duplex structures. Following the definition given in [1], the *NOESY graph* is an undirected graph situated on a plane. Each of its vertices represents one cross-peak from a corresponding NOESY spectrum, thus, a number of vertices equals the number of cross-peaks in the selected region of the spectrum. Vertices are weighted due to their correlation to intra- (weight 1) or internucleotide (weight 0) signals. An edge in the NOESY graph is added between every two collinear vertices having different weights. Thus, the NOESY graph contains only horizontal and vertical edges. Having such a graph model, we have defined the *NOE pathway* in the graph [1]. It alternately passes through vertices with different weights, starting from a vertex with weight 1, and changes the direction (horizontal / vertical) at each vertex. Moreover, the path does not contain collinear edges and satisfies the conditions of a Hamiltonian path. Thus, the problem of finding the NOE path in the NOESY graph is equivalent to a version of a Hamiltonian path problem, which we have named *Manhattan Hamiltonian path*. Figure 8 shows a fragment of an example NOESY spectrum and the corresponding NOESY graph.

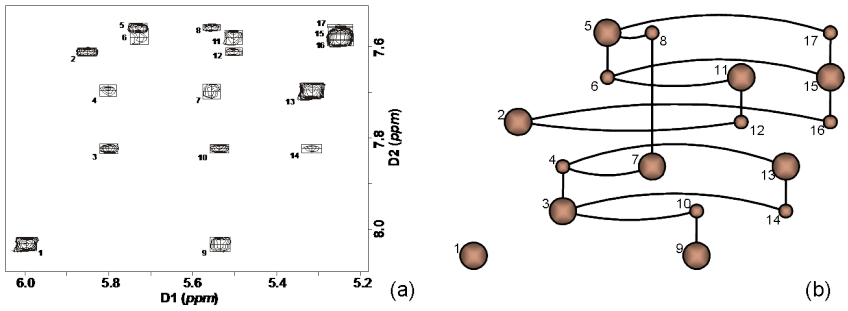


Figure 8: NOESY spectrum (a) and the corresponding NOESY graph (b)

Let us observe that the set of vertices of the NOESY graph is divided into two separate subsets: subset of vertices with weight 1 and subset of vertices with weight 0. By the definition, there are no edges that connect the vertices from the same subset. Thus, the NOESY graph is indeed a bigraph and the NOE pathway is a Hamiltonian path in a NOESY bigraph. Such an interpretation simplifies the problem if we analyse the spectrum which contains a lot of overlapping cross-peaks. Then, it is hardly possible to construct a pathway which does not contain collinear edges and thus, this condition from the NOE path definition can be ignored. However, it is important to remember that the edges of bigraph represent horizontal or vertical conections between the cross-peaks and the path crosses them alternately. Thus, the information about edge direction must be saved in the graph structure. Labelling or coloring of the edges allows to store this information. Summing up, we can provide the following definition of the *NOESY bigraph*: let G=(V,E) be 2-edge-colored bigraph representing 2D NOESY spectrum of RNA molecule. V is a set of vertices divided into two separate subsets, Va and Vb, where Va contains vertices representing intranucleotide correlation signals, Vb contains vertices representing internucleotide correlation signals. A number of vertices in V equals a number of cross-peaks in the corresponding NOESY spectrum. E is a set of edges, where every edge connects a vertex from subset Va with a vertex from subset Vb. Every edge is colored depending on the direction of a corresponding connectivity in the spectrum: green edge corresponds to a vertical connectivity, while a black edge corresponds to a horizontal connectivity. Having the above formulation we can define the *NOE pathway* in the following manner. Let us consider the Hamiltonian pathway in the NOESY bigraph. We will call it the NOE pathway if the following conditions are satisfied: the pathway alternately goes through the vertices from subsets Va and Vb, passing by turns through green and black edges. Figure 9 shows an example of NOESY bipartite graph (a) representing 2D spectrum from Figure 8.

A collection of NMR experiments performed for different RNA molecules have shown that in the real case the differentiation between intra- and internucleotide correlation signals is hard. Thus, it is almost impossible to create two separate subsets of vertices while modelling the problem according to

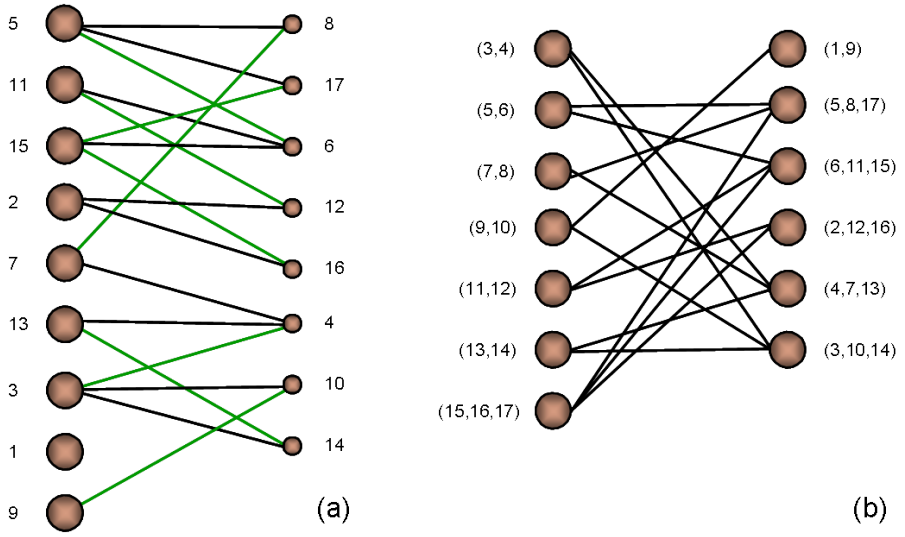


Figure 9: NOESY bipartite graph (a) and line bigraph (b)

the rules defining NOESY graph or NOESY bipartite graph. This inspired us to look for yet another representation which does not base on the separation of inter- and intranucleotide signals. The third graph model has been designed as a *NOESY line bigraph* G=(V,E). Vertices from set V represent the atoms which generate the observed NMR signals, thus, every vertex reflects possible connectivities that can be drawn between collinear cross-peaks for a given value of chemical shift. Since the connectivities are either vertical or horizontal, the set V is divided into two subsets, Vv and Vh, where Vv represents all the vertical connectivities in the corresponding 2D NOESY spectrum and Vh represents all the horizontal connectivities. Every edge from set E represents one cross-peak from the spectrum. The *NOE pathway* in the line bipartite graph passes alternately through the vertices from subset Vv and Vb, thus, satisfying the condition of perpendicular transitions between the cross-peaks in the corresponding NMR spectrum. An example line bigraph has been presented in Figure 9b. Numbers given for each vertex inform about all the collinear cross-peaks that belong to the connectivity represented by the vertex.

All of the presented models have been used to solve the sequential assign-

ment problem in 2D NMR spectra of short RNA molecules. However, 2D spectra are not sufficient to determine more complex structures. Nowadays, we can observe a big advance in the size of studied molecules. Obviously, the number of correlation signals recorded during NMR experiment grows with the molecule size, what results in obtaining a huge percent of overlapping cross-peaks. Their high density disrupts or even disables resonance signal identification on the basis of two dimensional experiments. A step towards three dimensions is the most evident solution to this problem. Let us then focus on a novel approach to an analysis of 3D spectra of more complex RNA molecules. From among many different 3D NMR experiments, three are used for sequential assignment: HCP, HSQC-NOESY, and NOESY-NOESY [6]. Each of these types serves an analysis of other correlation signals. However, the procedure of assignment is common for all. It starts from the identification of the sequence-specific connectivity pathway representing magnetization transfer between the selected nuclei of the analyzed molecule. A single connection in the pathway links either two cross-peaks having one common coordinate or two cross-peaks having two common coordinates, depending on a type of interactions analysed (homo- or heteronuclear). To formulate the problem in 3D let us denote by dFi(a,b) the direction of an edge between cross-peaks a, and b, having different coordinates in Fi dimension, and denote by dFiFj(a,b) the direction of an edge between cross-peaks a, and b, which differ in dimensions Fi and Fj. Now, we can define a *3D spectral graph*, representing an assignment problem in 3D NMR spectrum. Let G=(V,E) be an undirected graph satisfying the following conditions: every vertex from set V represents one cross-peak from 3D NMR spectrum (a number of vertices equals a number of cross-peaks), every edge from set E is assigned a label l(e(m,n))=0,1,2,3,4,5, where l(e(m,n)=0 if dF1(m,n), l(e(m,n)=1 if dF2(m,n), l(e(m,n)=2 if dF3(m,n), l(e(m,n)=3 if dF2F3(m,n), l(e(m,n)=4 if dF2F1(m,n), l(e(m,n)=5 if dF3F1(m,n), a number of edges in graph G equals all possible connections that can be drafted in the spectrum. Now we can formulate an *assignment pathway* in the 3D spectral graph. The pathway is a sequence of noncollinear edges in which every vertex and every edge occurs at most once. It is constructed according to one of the following principles: if we consider homonuclear interactions we construct a pathway from the edges

labelled 0,1,2 so that l(j)<>l (j+1) and l(j)<>l(j+2); if we consider heteronuclear interactions edges in the pathway follow the rule (l(j) mod 3)=(l(j+1) mod 3). Let us notice that according to the above definitions we obtain 6-edge-colored graph representing a 3D NMR spectrum. In this graph we reconstruct either 2- or 3-colored pathway which reflects magnetization transfer between molecule atoms. Figure 10 presents an example of 3D NMR spectrum projected on a plane XY and the corresponding spectral graph.

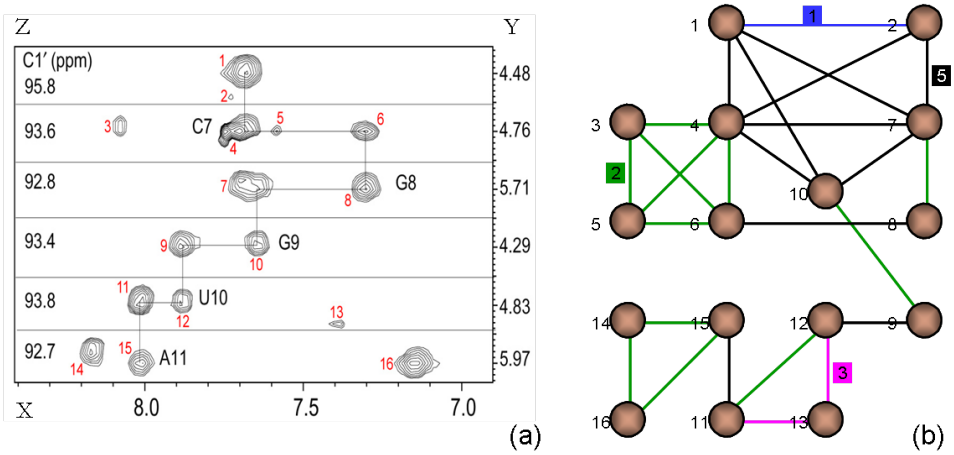


Figure 10: 3D spectrum projected on XY plane (a) and the corresponding graph (b)

3 Conclusion

In this paper, we have presented graphs used to model the problem of sequential resonance assignment in NMR spectra recorded for RNA molecules. On the basis of the first model, proposed for the ideal case in 2D, we have proved strong NP-hardness of the problem of the NOE path reconstruction in the NOESY graph [1]. All of the models have been used by the algorithms for automatic reconstruction of transfer pathways in 2D NMR spectra [3]. We have also described the graph theoretical model for the assignment problem in three dimensional space [6]. The model has been used by the enumera-

tive algorithm processing 3D NMR spectra and is being tested on the set of experimental data.

References

[1] R.W. Adamiak, J. Blazewicz, P. Formanowicz, Z. Gdaniec, M. Kasprzak, M. Popenda, M. Szachniuk. An algorithm for an automatic NOE pathways analysis in 2D NMR spectra of RNA duplexes. *Journal of Computational Biology*, 11/1:163–180, 2004.

[2] J. Blazewicz, A. Hertz, D. Kobler, D. de Werra. On some properties of DNA graphs. *Discrete Applied Mathematics*, 98:1–19 1999.

[3] J. Blazewicz, M. Szachniuk, A. Wojtowicz. RNA tertiary structure determination: NOE pathways construction by tabu search. *Bioinformatics*, 21/10:2356–2361 2005.

[4] H.H. Gan, S. Pasquali, T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Research*, 31/11:2926–2943 2003.

[5] P. Micikievicius, N. Deo. Exploring topological properties of NMR graphs. *BIBE*, 1304–1307 2007.

[6] M. Szachniuk, M. Popenda, R.W. Adamiak, J. Blazewicz. An assignment walk through 3D NMR spectrum. *IEEE CIBCB Proceedings*, 215–219 2009.

[7] S. Vishveshwara, K.V. Brinda, N. Kannan. Protein structure: insights from graph theory. *Journal of Theoretical and Computational Chemistry*, 1/1:187–212 2002.

ProDomAn - Protein Domains Analysis platform

Piotr Lukasiak*[2,3]*, Maciej Antczak*[2]*, Arkadiusz Hoffa*[2]*,

***Wojciech Biniecki*[4]*, Michal Wojciechowski*[1,4]

1 Introduction

Understanding the nature of proteins, especially knowledge about their particular functions as well as about their domains, requires crystallization of proteins. This process is very time-consuming because the current state of art of protein knowledge delivers almost no information how to crystallize new protein, and usually such process needs to be repeated many times before finding proper combination of parameters of crystallization environment.

2 ProDomAn

ProDomAn is a web-based platform developed to introduce tools for prediction of proteins functions and domain identification. This platform has been integrated with existing web application called WebMobis. Algorithms used for bioinformatics computations have been designed and implemented

[1]E-mail: michal.t.wojciechowski@gmail.com

[2]Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60-965 Poznan, Poland

[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, ul. Z. Noskowskiego 12/14, 61-704 Poznan, Poland

[4]Institute of Computing Science, Poznan University of Technology student

to simplify the process of proteins structures prediction and analysis. ProDomAn is composed of two algorithms (DomAn2 used to predict domains recognition, and BSPred used to predict proteins functions). Both algorithms have been implemented as separate applications in order to reduce the amount of queries sent to the WebMobis application's database and to assure that computations are independent from the application's breakdown.

3 DomAn2 algorithm

The DomAn2 is the algorithm designed to predict domain of proteins boundaries. To provide this feature the algorithm uses domain's patterns (as a domain's pattern one can understand a pair of aminoacid's sequences located at the beginning and at the end of each segment of considered domain). All patterns are stored in database designed specially for the DomAn2 algorithm. Patterns are obtained from five domain's classification databases: CATH, Conserved Domain Database (NCBI), Dali Domain Dictionary, Pfam and SCOP. Conserved Domain Database is composed of many sequence alignments for old domains and full-length proteins. The Pfam database is a collection of protein domains families. It contains only a collection of multiple sequence alignments and hidden Markov models covering many protein's domains with unknown structure and thus none of them can be found in Protein Data Bank (PDB). Other databases that have been used contain detailed and comprehensive description of the structural classification, hierarchy and evolutionary relationships among all known proteins. All databases are monthly checked for updates and whenever it is possible new patterns are added to algorithm's database. Currently it stores over 74, 500 000 patterns.
The DomAn2 algorithm is composed of three main stages. In the first stage the algorithm tries to match domain patterns from database to every input protein sequence sent to the ProDomAn platform. For one-segment domain case, pattern is considered to be matched successfully only if both subsequences that belongs to pattern have been found in considered protein and simultaneously the first subsequence of the pattern is followed by the last subsequence of the same pattern. In discontinuous domain case (when

the domain is composed of more than one segment) all patterns have to be marked as matched to the input sequence and they have to follow original segments order. In next stage, algorithm predicts where domains can be found in analyzed protein. This stage is composed of domains that ensure the biggest covering of input sequence. If there is no matched pattern the sequence is considered as one continuous domain covering the whole protein sequence.

The DomAn2 algorithm took part in eighth edition of CASP Experiment

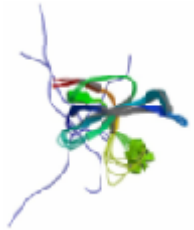
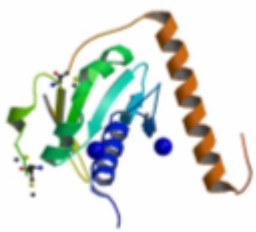
		
Protein: CA13A Organism: Homo sapiens Residues: 283	Protein: SaR32 Organism: Streptococcus galactiae Residues: 128	Protein: SPO1766 a protein of unknown function Organism: Silicibacter pomeroyi Residues: 150

Table 1: Domain divisions best predicted by DomAn2 algorithm [1].

(Critical Assessment of Structures Prediction). In this experiment one hundred and twenty one unknown proteins have been released. DomAn2 predicted domains correctly for 42 proteins (for 3 of them DomAn2 was the best of all).

4 BSPred algorithm

The second algorithm implemented for ProDomAn platform was BSPred. This algorithm provides ability to predict protein's functions. It predicts binding sites between ligands and amino acids in protein. Based on ligands contexts stored in specially dedicated database, algorithm tries to match ligands patterns to the input sequence. The pattern must be matched in the same way as in the DomAn2 algorithm. If any pattern can not be matched to the sequence then no binding site can be determined and no function can be predicted. Every predicted binding site is being adjusted to the proteins domains. If such adjustment is possible, it is considered that binding site has a function of matching domain.

5 Conclusion

Currently the ProDomAn platform is available in WebMobis application: `http://www.webmobis.cs.put.poznan.pl`

References

[1] *Research Collaboratory for Structural Bioinformatics (RCSB).* `www.rcsb.org/pdb/`, 2009.

CompuVac – development and standardized evaluation of novel genetic vaccines

***Piotr Lukasiak*[1,2,3], *Jacek Blazewicz*[2,3], *David Klatzmann*[3]**

Vaccination is the generic term for immunization procedures. Immunization is a procedure whereby living or nonliving materials are introduced into the body in order to stimulate the defence system of the host, directed against micororganisms, namely the lymphoid tissues, with a view to rendering a person or animal immune to a disease. This procedure is designed to increase concentrations of antibodies and/or effector T-cells, which are reactive against infection. These antibodies and T-cells recognize foreign proteins (so called antigen or immunogen) from microorganisms. Historically, vaccination was born in 1796 when Edward Jenner, an English physician who, during his practice in the countryside, noticed that farmers exposed to infected materials from cows did not develop smallpox but acquired immunity to the disease. Jenner decided to use the material derived from the bovine (vaccinus) lesions to vaccinate a boy (James Phipps), and showed that the patient was immune to a subsequent challenge with smallpox. The scientific approach to vaccination came only a century later, when Louis Pasteur introduced the concept that infectious diseases were caused by microor-

[1]E-mail: piotr.lukasiak@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
[4]Universite Pierre et Marie Curie, 7087 - UPMC-CNRS, Paris, France

ganisms and discovered that they could be attenuated by growing them under adverse conditions. Using this empirical approach he developed the first live-attenuated bacterial and viral vaccines (chicken cholera bacillus and rabies, respectively). Large-scale vaccination came only following the discovery by Glenny and Hopkins in 1923 and Ramon in 1924 of safe and reproducible ways to inactivate toxins and pathogens by the use of formaldehyde, and of the stable attenuation of pathogens by serial passage in vitro. These simple, basic technologies were the only means used for vaccine development for most of the 20th century. Later, the progress in recombinant DNA and conjugation technology in the 1970s and onwards, allowed the development of vaccines against those infectious agents that were difficult or impossible to grow in the laboratory. The new generations of vaccines are:

1. Peptide vaccines
2. DNA vaccines
3. Vector vaccinesVirus
    - Antigen Delivery Systems
    - Bacterial Antigen Delivery Systems
4. Virus Like Particles

Peptide vaccines are developed using epitopes of e.g. a surface protein from a microorganism as vaccine. These antigenic determinants are recognized by antibodies and T cells, but are not so immunogenic and need to be used with an adjuvant. DNA vaccines are developed by insertion of the gene of the antigen into a plasmid vector. The vaccine is prepared with the naked DNA plasmid and an adjuvant. Genes are often expressed in skeletal muscle cells or adipocytes, where they facilitate an immune response. Vector vaccines are derived from poxvirus, adenovirus, retrovirus, alphavirus and herpes virus. They are developed by replacement of viral replication genes with the gene encoding the antigen. After vector entry the antigen is synthesized. Antigens can be processed into peptides, which are presented at the cell surface by MHC complexes; and this induces as cellular immune response. Antigens

that are and not processed into peptides can be secreted and induce an antibody response. Virus-Like Particles (VLPs) as immunogens - most of the currently used antiviral vaccines are based on homologous inactivated or attenuated viral particles, which suggests that such particles are highly immunogenic. This would indicate that presenting an antigen into/onto particles should be an efficient way of generating an immune response. VLPs consist of one or more viral coat proteins that assemble into particles and mimic the overall structure of vaccine particles without the requirement of containing infectious genetic material. VLPs have been produced for more than 30 different viruses. In addition to the use of VLPs as a direct immunogens substituting for the specific virus, VLPs can also be used as vehicles for the delivery of antigens, or antigen epitopes from other microoganisms. CompuVac [1] is a project financed by the European Commission, which involved 18 partners worldwide. CompuVac's main objectives are to setup a standardised approach for the rational development of genetic vaccines and to apply this methodology to the development of vaccines against the hepatitis C virus. Recombinant viral vectors and virus-like particles are considered the most promising vehicles to deliver antigens in prophylactic and therapeutic vaccines against infectious diseases. Several potential vaccine designs exist but their cost-effective development cruelly lacks a standardized evaluation system. On these grounds, CompuVac is devoted to *(i)* rational development of a novel platform of genetic vaccines and *(ii)* standardization of vaccine evaluation. CompuVac assembles a platform of viral vectors and virus-like particles that are among today's most promising vaccine candidates:

- The viral / bacterial vector platform comprises vectors derived from adenovirus, herpes virus, measle virus, vaccinia virus Ankara (MVA) and Mycobacterium bovis Bacille Calmette-Guerin (BCG).

- The VLP vector platform comprises VLP derived from murine retrovirus, polyoma virus, bacteriophage and Hepatitis B Virus. Retroviral-derived VLPs devoid of any genomes produced by plasmids, or "plasmoVLP", and pure polyepitope tags, without any VLP carrier, was also developed.

CompuVac recognizes the lack of uniform means for side-by-side qualitative and quantitative vaccine evaluation and will thus standardize the evaluation of vaccine efficacy and safety by using "gold standard" tools, molecular and cellular methods in virology and immunology, and algorithms based on genomic and proteomic information. "Gold standard" algorithms for intelligent interpretation of vaccine efficacy and safety will be built into CompuVac's interactive Genetic Vaccine Decision Support system (GeVaDSs) [2]. As end products, vector platform and "gold standard" tools, methods and algorithms will be available to the scientific and industrial communities as a toolbox and interactive database which standardized nature should contribute to cost-effective development of novel vaccines. The large availability described above to create different types of novel vaccines creates a need for standardized evaluation. For this purpose CompuVac has assembled a platform of viral vectors and VLPs that are among today's most promising vaccine candidates. These vaccines are described in more detail on the vectors platform section. These vaccine candidates are evaluated by using standardized protocols and tools and GeVaDSs is built-up in order to generate *(i)* vector classification according to the type of induced immune response and quality, *(ii)* vector combination counsel for prime-boost immunizations, and *(iii)* vector molecular signature according to genomic analysis. Knowledge assembled from these studies will be applied to the development of vaccines against Hepatitis C Virus.
CompuVac mainly focused on:

1. *Recombinant defective viral/bacterial vectors*

As it was said viral vectors, alone or in combination in prime-boost strategies, are considered to belong to the most promising vehicles to deliver antigens for prophylactic or therapeutic vaccination against infectious diseases. Reasons are extremely efficient cell transduction in vivo and strong adjuvant effects provided by viral functions. Although there have been many studies at the preclinical or clinical level, in which viral vectors have been used for vaccination, surprisingly, there have been no studies that allow side-by-side comparison of viral vectors in a systematic and standardized way. The main objective is to generate different viral vectors expressing LCMV and VSV model antigens such as to characterize resulting immune responses

against these antigens in standardized animal models. The best vectors or vector combinations will be used to generate constructs with HCV envelope as antigen. The vectors developed are derived from the following viruses and bacteria: adenovirus, herpes virus, measle virus, vaccinia virus Ankara (MVA) and Mycobacterium bovis Bacille Calmette-Guerin (BCG). In order to allow side by side comparison of immune responses, the vectors are designed to express the model and therapeutic antigen, respectively, in exactly the same configuration. It was decided to use the hCMV promoter to control gene expression. This is a widely used and very strong promoter that is ubiquitously active in different tissue. It was also decided to use the SV40 polyadenylation signal for transcript termination.

2. Inert Virus-like particles

The absolute benefit of working with a wide variety of inert particle vaccines is avoiding the necessity of viral vectors and their genes, thereby eliminating all risks connected with this kind of therapy. Moreover, one additional advantage when working with non-human inert particle vaccines is that one avoids possible pre-existing immunity in humans to the vectors. This pre-existing immunity could otherwise present a problem by neutralizing the vaccine vector and in this way abrogating the expected immune response. The main objective is to generate different VLPs expressing LCMV and VSV model antigens and use them as such to characterize the resulting immune responses against these antigens in standardized animal models. The best VLPs or VLPs combinations are used to generate constructs with HCV envelope as antigen. The actual platform of inert particle vaccines consists of VLP derived from murine retrovirus, polyoma virus, bacteriophage and Hepatitis B Virus. Retroviral-derived VLPs devoid of any genomes produced by plasmids, or "plasmoVLP", and pure polyepitope tags, without any VLP carrier, was developed.

3. Standardized immunological evaluation of vaccine efficiency

The main objective is to establish a standardized vaccine test system, which allows a stringent comparison of vaccination efficacy obtained by different laboratories at different time-points. For this purpose a toolbox of standardized assays and reagents will be created and provided. It is serving as a pre-read-out to select efficient vaccines against HCV and will identify immun-

odominant epitopes as targets for neutralizing antibodies to HCV. The efficacy of different gene expressing vaccines and inert particle vaccines will be defined in terms of induction of cytotoxic T cell responses and humoral responses. For monitoring the cellular immune responses one has chosen as a gold standard antigen a short peptide sequence; gp33-41 from the LCMV envelope. The main driving reasons for this choice are:

- Cytotoxic T lymphocyte (CTL) immune responses induced by this peptide are very well characterized and can be quantitatively measured
- The T cell receptor (TCR) recognizing this peptide in the H-2b background has been cloned
- Transgenic mice expressing this TCR have been generated
- Tetramers detecting this TCR have been generated

Excellent infectious models exist that permit to evaluate and correlate the CTL responses with protection from infectious challenge For monitoring the humoral responses, one has chosen as a gold standard antigen the envelope proteins of the Vesicular Stomatitis Virus serotype Indiana (VSV). The main driving reasons for this choice are:

- Standardized quantitative assays exist to measure the humoral responses against these proteins
- The tools to assay immune responses to this virus, including neutralization assays with live virus are available
- VSV envelope protein can easily be pseudotyped onto retroviruses, and thus suitable for developing neutralization assay with defective chimearic viruses

This vaccine characterization will also involve epitope mapping and microarray expression profiling which reflect the molecular signature of the vaccine. The core of the research within Compuvac is GeVaDSs - system created for efficient storage, integration, retrieval of data, and moreover, for intelligent association and interpretation of data, and thus for knowledge-based generation of testable hypotheses of immune responses to antigens developed in

CompuVac consortium One of the main feature of GeVaDSs is a standardization of the way immunological experiments are proceeded and desribed. GeVaDSs allows not only for confident comparison of experiments and analysis of the results, but also delivers detailed description of experiments (immunization protocol, results, analysis) as well as generates all data sheet templates needed to upload and store data during experiments. Using GeVaDSs it is easy to find differences and strong points of different vectors, antigens or immunization protocols. GeVaDSs is divided into two parts from functional point of view:

a) interactive database section that include

- results generated using the tool box standardized protocols
- assumed standard algorithms to comparatively assess vaccines to be developed with standardized methodology developed by consortium
- results of upcoming preclinical and clinical European studies.

b) support section that include

- algorithms for the intelligent comparison of new vectors to previously analyzed ones).
- vector evaluation
- error prediction
- statistical analysis From scientific point of view three main sections can be found in GeVaDSs:
- section responsible for T cell data,
- section responsible for B cell data
- and section responsible for molecular signature. Current version of GeVaDSs gives a following functionality (between others): vector categorization tree
- immunization protocol definition
- uploading and analysis of T cell experiments as well as comparison between T cell experiments uploading and analysis of individual,

group and experiment results of B cell experiments as well as comparison between B cell experiments

- uploading and analysis of molecular signature results (microarray analysis), gene filters definition, results analysis as well as comparison between molecular signature experiments
- all data can be analyzed and presented in various formats and graphical forms.

CompuVac aims to generate and make available to the scientific community a *"tool box"* and an *"interactive database"* allowing for the comparative assessment of future vaccines to be developed with our gold standards. We believe that this should have a significant impact on vaccine development, and notably for those vaccines requiring prime/boost immunizations.

References

[1] CompuVac: `http://www.compuvac.org`

[2] GeVaDSs: `http://gevads.cs.put.poznan.pl`

A certain model of HCV virus infection

Szymon Wasik[1,2]***, Paulina Jackowiak***[3]***, Jacek Krawczyk***[4]***,***

Pawel Kedziora[2]***, Piotr Formanowicz***[2,3]***, Marek Figlerowicz***[3]***,***

Jacek Blazewicz[2,3]

Acknowledgements: This study was partially supported by the Polish Government through grants N301 019 31/0483 and N519 314635 from the Ministry of Science and Higher Education. In addition P. Jackowiak was supported by the scholarship provided by the President of the Polish Academy of Sciences.

1 Introduction

Hepatitis C virus is one of the most prevalent human pathogens. According to WHO about 1.5% of the world population is infected [7]. It can persist in the host for the long time and causes chronic infections that can lead to liver fibrosis, cirrhosis and hepatocellular carcinoma [1, 6]. Currently no effective vaccine or way of treatment is known. Without treatment only about 10% of patients recover. Current treatment scheme increases this value to 40% [5].

The good mathematical model of the HCV infection can help in finding

[1]E-mail: szymon.wasik@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
[4]Faculty of Commerce and Administration, Victoria University of Wellington, Wellington, New Zealand

an effective treatment method. The aim of this work is to propose a model and its application for early stage assessment of patients infected with HCV. To evaluate the model we used data gathered during the case study in Poznan hospitals. These are an RNA level, phylogenetic trees and a mean Hamming distance.

The standard therapy lasts 72 weeks. The level of the HCV viral RNA accumulation is determined for each patient at the beginning of the therapy and then after 24, 48 and 72 weeks. These weeks will be referred as T0, T24, T48 and T72. Patients with a detectable amount of HCV RNA in T24 are excluded from further therapy and qualified as a patient with *no response*. The others are treated till T48. At T72, patients response to the treatment is assessed and qualified as *sustained response* if viral RNA is not detectable or *transient response* otherwise. Some time ago during the research in Poznan [2, 3] some factors were checked if they can explain the type of the response. During the research it was determined that number of viral variants in the blood could not explain the type of the response but using mean Hamming distance as a measure of HCV genetic diversity and phylogenetic trees it was possible to precisely predict the response.

2 Statistical analysis

The only problem with the case study described above was that it was based on only 15 test cases and getting more data about Hamming distance and phylogenetic trees is difficult and expensive process. To deal with this problem we checked if there exists a correlation between Hamming distance and the viral RNA level. The result is presented in Figure 11. The R^2 coefficient for this regression is equal to 0.39. This means that about 40% of variability in the mean Hamming distance can be explained by changes in the RNA level. It can be considered as rather large so we are confident that we can use patients' RNA levels as a proxy for HCV genetic diversity and hence forecast patient curability on the basis of the former. For the RNA level we currently have data about over 100 test cases and gathering additional data is much simpler. As an alternative for the RNA levels we have also examined Hamming distance dependence on the ALT level but the R^2 constant was much

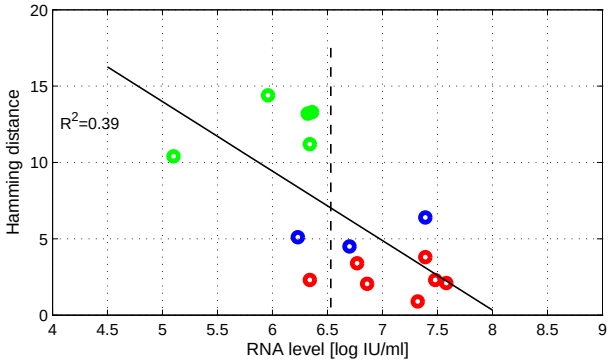


Figure 11: Linear regression between the virus RNA level and the Hamming distance of sequences at T0. Green colour represents sustained response, blue transient response and red no response.

smaller. To verify the analysis we have applied the F test for the model and the t tests for the coefficients. These tests have confirmed significance of the model.

We have also checked the distribution of data about the RNA level at weeks 0, 24, 48 and 72. The Lillifors test confirmed that the distribution of data is normal and $\mu = 5.84, \sigma = 0.88$ at the beginning of the treatment.

3 Transition states and matrices

We have divided patients into 3 groups. Group N (no) with the undetectable RNA level, group M (medium) with the RNA level below $6.53 \frac{\log \text{IU}}{\text{ml}}$ and group H with the viral RNA level above this value. The $6.53 \frac{\log \text{IU}}{\text{ml}}$ threshold was chosen to separate the group of patients with the sustained response from the first patient with the transient response. It is also close to $\mu + \sigma$. Based on

these states we defined transition probability (stochastic) matrix as follows:

$$\mathcal{T}_{i,j} = \begin{bmatrix} p_{N,N}^{(i,j)} & p_{N,M}^{(i,j)} & p_{N,H}^{(i,j)} \\ p_{M,N}^{(i,j)} & p_{M,M}^{(i,j)} & p_{M,H}^{(i,j)} \\ p_{H,N}^{(i,j)} & p_{H,M}^{(i,j)} & p_{H,H}^{(i,j)} \end{bmatrix} . \tag{1}$$

Here, i and j are indices of weeks between which a transition occurs ($i, j \in \{T0, T24, T48, T72\}$) and $p_{g,h}^{(i,j)}$ is a transition probability between two groups $g, h \in \{N, M, H\}$ and weeks i and j. We also defined a vector containing numbers of patients at the beginning of week i as:

$$P_i = \begin{bmatrix} P_{i,N} & P_{i,M} & P_{i,H} \end{bmatrix} \tag{2}$$

where $P_{i,g}$ denotes the number of patients in group g at week i. Using Equations 1 and 2 we can calculate the number of patients in week j as:

$$P_j = P_i \cdot \mathcal{T}_{i,j} . \tag{3}$$

4 Therapy efficiency

Because the currently used therapy has significant side effects it would be beneficial if only those patients are treated who have a high probability of developing a sustained response. This would improve the treatment efficiency because non-curable patients would be spared suffering. We define the therapeutic efficiency ratio as follows:

$$\varepsilon = \frac{\texttt{total cured}}{\texttt{total treated}} , \quad \varepsilon \in [0,\, 1] . \tag{4}$$

In our research we checked how the therapy efficiency would change when only patient below some viral RNA level are qualified for the treatment. We assumed that only patients with the RNA level below some qualification level M_{max} are treated where M_{max} is the viral RNA level that separates groups M and H. We have used two algorithms to predict therapy efficiency for different values of M_{max}. First of them is following:

1. Calculate transition matrices $\mathcal{T}$ using a new value of M_{max}.

2. Replace the last row of $\mathcal{T}_{0,24}$ with the last row of $\mathcal{T}_{48,72}$.
3. Calculate the number of patients P_{72} at T72 using Equation 3.
4. Calculate the value of $\varepsilon(M_{max})$.

The second step of this algorithm is based on the observation that $\mathcal{T}_{48,72}$ can represent transitions when patients are not treated. If we replace the last row in $\mathcal{T}_{0,24}$ with the last row of $\mathcal{T}_{48,72}$ we will receive transitions when only patients from group M are treated. This algorithm uses transition matrices to estimate the value of ε. We compared it to the algorithm that does not use these matrices and calculates values of ε using the raw data. This algorithm is following:

1. Select only these patients who have the level of RNA at T0 lower or equal to M_{max}. Let P_{All} denotes the number of these patients.
2. Calculate the number of patients who had sustained response in the selected group of patients. Let P_{SR} denotes the number of these patients.
3. Calculate the value of ε using the following formula:

$$\varepsilon = \frac{P_{SR}}{P_{All}} \tag{5}$$

The results are presented in Figure 12. The first algorithm uses results of

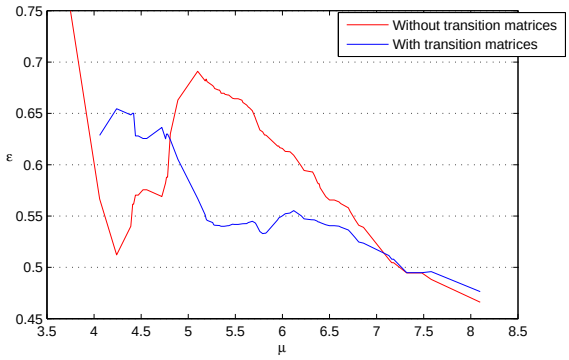


Figure 12: The value of therapy efficiency for different values of M_{max}.

statistical analysis to estimate the value of ε instead of non-processed data and that is why it is more noise-resistant. It also utilizes more information than the simple, second algorithm. That is why it estimates function without any significant local minimum which should not occur. We can also observe that increasing M_{max} decreases the value of ε. It is consistent with other reports (for example [4]). It can be noticed that after exceeding the RNA level of $M_{max} = 5.25\frac{\log \mathrm{IU}}{\mathrm{ml}}$ the efficiency of the therapy modelled with transition matrices decreases significantly and the efficiency of the therapy modelled without transition matrices stays at the constant level so it can be a good therapy qualification level.

5 Summary

In this article we used a linear regression analysis to test the dependency between the genetic variability of the virus and its RNA level in blood. R^2 coefficient for the regression was reasonably high which enabled us to use the level of viral RNA accumulation as a proxy for the genetic variability. We then defined three patient groups separated by the different viral RNA levels. Next, we constructed matrices describing transitions between the groups. Finally we analysed the therapeutic efficiency using unprocessed data and results of the algorithm that uses transition matrices. Our results indicated that viral RNA level below $5.25\frac{\log \mathrm{IU}}{\mathrm{ml}}$ can be regarded as a threshold separating patients with high probability to develop a sustained response from the others.

References

[1] A. Alberti, L. Chemello, L. Benvegnu, Natural history of hepatitis C, *Journal of Hepatology*, Supplement 1, pages 17–24, 1999.

[2] M. Figlerowicz, P. Jackowiak, M. Alejska, N. Malinowska, A. Kowala-Piaskowska, P. Kedziora, P. Formanowicz, J. Blazewicz, M. Figlerowicz, Two types of viral quasispecies identified in children suffering from chronic hepatitis C, *Journal of Hepatology*, page S127, 2009.

[3] P. Kedziora, M. Figlerowicz, P. Formanowicz, M. Alejska, P. Jackowiak, N. Malinowska, A. Fratczak, J. Blazewicz, M. Figlerowicz, Computational methods in diagnostics of chronic hepatitis C, *Bulletin of the Polish Academy of Sci. Tech*, pages 273–281, 2005.

[4] J.Y.N. Lau, G.L. Davis, J. Kni®en, K.P. Qian, M.S. Urdea, M. Chan C. S. abd Mizokami, P.D. Neuwald, J.C. Wilber, SigniŻcance of serum hepatitis C virus RNA levels in chronic hepatitis C, *Lancet*, pages 1501–1504, 1993.

[5] J.G. McHutchison and T. Poynard, Combination therapy with interferon plus ribavirin for the initial treatment of chronic hepatitis C, *Seminars in Liver Disease*, Supplement 1, pages 57–65, 1999.

[6] L.B. Seeff, Natural history of chronic hepatitis C, *Hepatology*, pages S35-46, 2002.

[7] World Health Organization, Hepatitis C global prevalence (update), *Weekly Epidemiological Record*, pages 425–427, 1999.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

An application of hyperheuristics

Aleksandra Swiercz[1,2,3]

Acknowledgements: The work has been partially supported by the Polish Ministry of Science and Higher Education grant NN 5193 14635.

1 Introduction

Hyperheuristic is a new searching technology, which goal is to raise the level of generality at which optimization system can operate. Heuristics and metaheuristics use a lot of domain knowledge and are very problem-specific. The implementation of such methods is costly, however the results are usually close to optimum. Hyperheuristics are developed for the wide range of problem domains. They are working at a higher level in comparison to typical methods developed for the optimization problems. Instead of searching in the solution space, a hyperheuristic employs low-level (meta)heuristics, which are dedicated to the problem, to search the solution space. It chooses in intelligent way which heuristic should be applied at each moment. Hyperheuristic can be applied to many different problems only by changing the low-level heuristics. They are not dependent on the domain knowledge. The only information which is transmitted between low-level heuristics and

[1]E-mail: aswiercz@cs.put.poznan.pl
[2]Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704, Poznań, Poland

a hyperheuristic is the status of the improvement of the obtained solution. The hyperheuristic has to learn the way in which to choose an appropriate heuristic. A hyperheuristic framework can be found in Figure 13. As a result of the search process one can obtain the solution which is "good enough - soon enough and cheap enough".

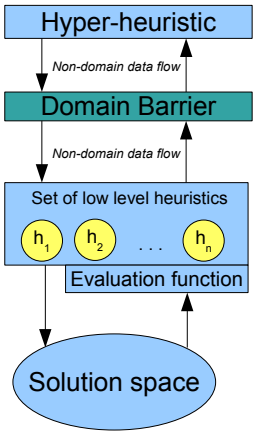


Figure 13: A hyperheuristic framework

The term hyperheuristic is quite new, however the concept of the method origins from the field of Artificial Intelligence on automated planning systems. The earliest of such systems tried to find a series of actions that achieve a given goal. Usually the goal was to reduce the difference between the current state of the world and the desired world. An example of the such system was the DART logistical planning system, which was used in the Gulf War [5]. More sophisticated example could be a system for planning the schedule of communication between satellites and ground stations - LR-26 scheduler [6].

2 A hyperheuristic framework

A possible framework of a hyperheuristic could be:

- Define set H of low-level heuristics. Each heuristic can transform a

problem state (a solution) into a new problem state.

- Start with an initial solution S_0. The number of iterations t is equal to 0.
- Increment the number of iterations t. Check the heuristics from set H (some or all of them), select one of them and apply it to the problem state S_{t-1} tranforming this state to the new one, S_t.
- If the problem is solved, stop. Otherwise go to the previous step.

The set of heuristics is developed for the specific problem. Some of the heuristics should work randomly in order to introduce enough diversity into the search process, while the other aim to improve the solution. Each heuristic can transform a solution into another one. The improvement of the objective value for every transformed solution is evaluated by the function. A hyperheuristic could ask at each iteration t every heuristic from the set H. However, the computation can be very expensive, and variuos techniques are used in order to optimize the number of checked heuristics.

The way of checking and selecting low-level heuristics depends on the chosen hyperheuristic. A hyperheuristic could be for example a metaheuristic: tabu search, simulated annealing or genetic algorithm. Below the concept of the choice function method as a hyperheuristic is described.

3 The choice function method

The hyperheuristic based on the choice function method first appeard in [4]. The choice function method uses the ranking of low-level heuristics for selecting the most appropriate one at each iteration. The choice function is a weighted combination of three factors, and it is defined as $f(t, h_i) = \alpha f_1(t, h_i) + \beta f_2(t, h_i) + \gamma f_3(t, h_i)$, where t is the iteration number, h_i is the i-th heuristic from H, and α, β and γ are the weights of functions f_1, f_2, f_3. Functions f_1 and f_2 are designed to intensify the search, while function f_3 introduce the diversification.

At each iteration t the return value of function f is determined for every heuristic h_i which is checked by the hyperheuristic. Function $f_1(t, h_i)$ gives

the information about the recent efectiveness of heuristic h_i. The return value is equal to 0 if heuristic h_i has not been used yet, otherwise it is proportional to the improvement of the objective function value. Function $f_2(t, h_i)$ holds the information about the recent efectiveness of the pairs of low-level heuristics. It checks the improvement of the objective function value when heuristic h_i was applied after heuristic h_j. If heuristic h_i has not been used so far, the value of the function is equal to 0. The function $f_3(t, h_i)$ is the time span between the current iteration t and the last iteration that h_i was used. The return value of f_3 is very high when the heuristic has not been used for a long time.

The low-level heuristics are ranked according to function f. The hyperheuristic applies the selection mechanism (i.e. straight choice, ranked choice, decomp choice, roulette choice) in order to choose h_i which will be used at iteration t.

4 An application of hyperheuristic

The goal of the research is to check the behaviour of a few hyperheuristics (choice function method, tabu search and simulated annealing) to the sequencing by hybridization problem [7]. Many different algorithms (heuristics and metaheuristics) solving the problem have been proposed so far [1, 8, 3, 2]. The aim of this new approach is to compare different hyperheuristic methods working with the same set of low-level heuristics, rather then to beat the older algorithms. The created software for the hyperheuristic framework can be adapted later on for other problems, like for example traveling salesman problem, by changing only the set of low-level heuristics, which are specific to domain knowledge of the problem.

References

[1] J. Błażewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, and J. Węglarz. DNA sequencing with positive and negative errors. *J. Comput. Biol.*, 6:113–123, 1999.

[2] J. Błażewicz, C. Oğuz, A. Świercz, and J. Węglarz. DNA sequencing by hybridization via genetic search. *Oper. Res.*, 54(6):1185–1192, 2006.

[3] T.N. Bui and W.A Youssef. An enhanced genetic algorithm for DNA sequencing by hybridization with positive and negative errors. *Lect. Notes Comput. Sci.*, 3103:908–919, 2004.

[4] P.I. Cowling, G. Kendall, and E. Soubeiga. A hyperheuristic approach to scheduling a sales summit. In *PATAT '00: Selected papers from the Third International Conference on Practice and Theory of Automated Timetabling III*, page pp 176–190. Springer-Verlag, London, UK, 2001.

[5] S.E. Cross and E. Walker. Dart: applying knowledge-based planning and scheduling to crisis action planning. In M. Zweben and M.S. Fox, editors, *Intelligent Scheduling*. Morgan Kaufmann, 1994.

[6] J. Gratch, S. Chein, and G.de Jong. Learning search control knowledge for deep space network scheduling. In *Proceedings of the Tenth International Conference on Machine Learning*, pages pp 135–142. Lect. Notes Comput. Sci., 1993.

[7] E.M. Southern. United Kingdom Patent Application GB8810400, 1988.

[8] J-H. Zhang, L-Y. Wu, and X-S. Zhang. Reconstruction of DNA sequencing by hybridization. *Bioinformatics*, 19:14–21, 2003.

Introduction to microarray data analysis

Hanna Cwiek[1,2]***, Aleksandra Swiercz***[2,3]***, Piotr Gawron***[2]***,***

Jacek Blazewicz[2,3]

1 Introduction

Measuring gene expression levels with microarrays has become one of the basic tools of modern genomics. Authors participate in research based on microarray data analysis in which ralations between genes and their functions are studied. In this abstract a biochemical introduction to the area will be given. Authors will present a description of performed experiments and data aquiring methods and discuss the problems encountered during data analyses. Finally, authors' plans for further contribution will be presented.

2 Biological background

One of the greatest discoveries in the 20^{th} century was determining the form of genetic information container — a DNA double helix. The helix consists of two DNA strands twisted together. Each strand is a long sequence of four similar molecules that are repeated many times throughout a genome. These molecules, called nucleotides, differ only in nitrogenous bases and are abbreviated A (adenine), T (thymine), C (citosine), and G (guanine). They are

[1]E-mail: hanna.cwiek@cs.put.poznan.pl
[2]Institute of Computing Science, University of Technology, Poznan, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

organised in groups called genes, each of them having particular function in a living organism.

In order to recover their functions genes need to be transcripted, which means copied to complementary RNA strands. It is possible that functional RNA is the final product of gene transcription in case of non-protein coding genes. However, usually RNA fragments, whose each three consecutive bases encode one aminoacid, serve as instructions for creating proteins. It is the proteins that are directly responsible for effecting changes in organism's behaviour. The whole process of transforming information from genes into functional gene products is called *gene expression*.

A set of genes contained in a living cell determine what the cell can possibly do. When cell's behaviour is to be changed certain genes get activated and expressed. Thus by observing gene expression one can say what the cell is doing at the particular moment. Measuring changes in gene expression levels in various conditions helps to match genes to their functions in an organism. Measurement of the activity of genes (known as gene expression *profiling*) can be performed with the use of microarrays.

3 Microarrays

A DNA microarray is a matrix containing series of thousands of microscopic spots where short DNA sequences are attached. At each spot a number of copies of the same sequence are located. The sequences are single-stranded fragments of genes. The number of spots, their content and length of DNA sequences may differ for microarrays of different spieces and depending on microarray manufacturer. For species whose genome is known dedicated matrices can be created with spots exactly representing parts of genes from their genome. The organisms that are not known yet have to be analysed on matrices for other species, usually known to be closely related to them.

In a microarray experiment a specially prepared solution of cDNA (DNA complementary to RNA sequences from transcription) from the examined organism is placed on an array. Preparation comprises cloning of sequences and marking them with distinguishing substance, e.g., fluorescent or magnetic. Sequences from the solution bind to complementary strands on the

microarray and remain there while the unmatched ones are removed. Upon exciting the distinguishing substance, e.g., with laser or magnet, measurements of spots intensities which represent gene expression levels in the organism can be taken.

On one microarray one or two different solutions can be analysed, which corresponds to single and dual channel matrix respectively. In case of a dual experiment two different distinguishers need be used, e.g., cyanine dyes Cy3 and Cy5. Here only dual channel matrices are considered.

4 Result analysis

To determine gene functions gene expression profiles of organisms in variuos conditions have to be compared using statistical tests. If the difference of expression levels between the experiments is significant from statistical point of view, the analysed genes are likely to be responsible for organism's responding to certain treatment. Basic types of comparisions are: time *t* vs time *t+1*, samples treated vs untreated, mutated vs wild or diseased vs healthy.

There are two ways of comparing gene expression profiles: direct — used in case of single and dual channel matrices, where each profile can be compared directly to one another, and indirect — through a reference variant being always the same substance on one channel of a dual channel matrix with the other one changing.

During the analysis of results from microarray experiments a number of problems are to be faced:

- experiment errors, noise, false or inaccurate reads — the correctness of microarray spots have to be verified, some spots should be excluded from comparision,

- differences in intensities of distinguishers between channels in a dual channel experiment and differences in intensities of distinguishers between experiments — values of intensities need to be normalized, e.g., with respect to reference spots, according to a reference variant, equal

avarage spot intensity, equal intensity quantiles, normal intensity distribution, etc.,

- multiple spots on array containing different sequences corresponding to the same gene — measured intensities must be merged sensibly to obtain a reliable gene expression level,
- different set of genes located on matrices from different manufacturers — common subset of genes need to be found to enable comparisions between microarrays.

5 Research areas

The research that authors participate in is focused on gene expression profile analysis based on microarray experiment data. In the project two types of experiments are performed. First one is the analysis of gene expression of species whose genome is not known with use of related species' microarrays. For instance, tobacco is examined in respect of different types of tomato and potato microarray. On that basis conclusions about similarity of genes and their functions in the species are expected to be drawn. Second experiment aims at analysis of responses of a single species to different treatments. The current species of interest is *Arabidopsis Thaliana*, whose genome has already been well studied. Applied treatment, also known as stress, can be variuos chemical substances or environment changes. As an effect, functionally similar genes can be found and functional groups can be analysed.

Authors' responsibility in the project is microarray data analysis. It comprises preprocessing of acquired data, normalization, comparisions of different expression profiles and conducting statistical tests. At present the experiments are conducted with the use of the same type of microarrays and thus no problem of different gene subsets has appeared. Great part of the analyses is made with support of `R` environment where variety of useful methods are implemented.

Plans for the future include data visualisation, cluster analysis and comparision of different clustering algorithms. In the second stage of the project machine learning methods will be used to enable visual interpretation of ex-

periments' results. Selected clustering algorithms will be tested and compared, and results will be veryfied on acquired data in cooperation with biologists and biochemists.

References

[1] J.M. Berg, J.L. Tymoczko, L. Stryer. *Biochemistry, Fifth Edition : International Version.* W. H. Freeman, 2002.

[2] D.J. Lockhart, E.A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405:827–836, 2000.

[3] P.O. Brown, D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nat Genet*, 21:33–37, 1999.

[4] The Arabidopsis Genome Iniative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408:796–815, 2000.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

Extended targeted profiling to identify and quantify metabolites in 1-H NMR measurements

Frank-Michael Schleif[1,2]***, Thomas Riemer***[3]***, Uta Boerner***[3]***,***

Michael Cross[3]

Acknowledgments: This work was supported by the Federal Ministry of Education and Research under FZ:0313833 A, in the project NMR Metabolic Profiling of the Stem Cell Niche (MetaStem. We would like to thank Prof. Thomas Villmann (University of Appl. Sc. Mittweida) and Prof. P. Maass (University of Bremen) for fruitful discussions about sparse approximation and functional signal processing. Further we thank the whole MetaStem team for an effective collaboration.

1 Introduction

The profiling of metabolites is a key step in the modeling and analysis of biological systems. Recent approaches in stem cell biology focus on the modeling of the stem cell metabolism to reveal the underlying chemical pathways [4]. It is assumed that the analysis of the chemical pathways of stem

[1]E-mail: schleif@informatik.uni-leipzig.de
[2]University of Leipzig, AG-Computational Intelligence, Leipzig, Germany
[3]University of Leipzig, Interdisciplinary Center of Clinical Research, Leipzig, Germany

cells may lead to new approaches in e.g. the treatment of leukemia. To provide the underlying metabolic information different biochemical experiments and measurements employing mass spectrometric (MS) or nuclear magnetic resonance (NMR) devices are done. Here we focus on metabolic profiling studies of stem cell extracts using 1-H NMR measurements. The analysis of such measurements involves in general different preprocessing steps such as phase- and baseline-correction as well as smoothing and data reduction techniques [12, 3]. Details on the basic processing of NMR spectra used in this paper can be found in [10]. The profiling of metabolites in such measurements involves two main steps: the identification of the potentially unknown metabolite signatures in the signal and the estimation of the concentration of the metabolites with respect to the original biological samples. Multiple approaches have been published to solve this problem [1, 13, 14, 11] but are currently insufficient to be applied in a sufficiently automated way. This however is necessary to allow the high-throughput processing of such studies [9, 8]. Especially in stem cell research, the cells are cultivated and analyzed in a huge number of different experimental settings and the manual or only roughly automated analysis of the obtained spectra is very time consuming and error prone. We present an approach to improve this situation by a semi-automatic analysis of the spectra such that only minor, simple interaction steps are necessary and the processing of large data sets remains tractable. First a basic introduction in NMR spectra analysis is provided. Subsequently we review the recently published approach of Targeted Profiling (TP) [11] which will be extended in this work. In the remainder of the paper initial results on simulated and real life measurements are provided supporting the improved performance of the presented approach with respect to a manual expert analysis.

2 Metabolic profiling by Nuclear Magnetic Resonance Spectroscopy

Here we focus on the analysis of 1-H liquid NMR spectra obtained from stem cell extracts, however the approach is also transferable to labeled NMR experiments such as 13-C labeled samples. Subsequently it is further assumed

that the chemical preprocessing and experimental design follows roughly the guidelines given in [11]. The obtained spectra are high dimensional, here we consider ≈ 30000 measurement points per spectrum with a resolution of 700.153 MHz, although the approach is more generic and also applicable in case of measurements with lower resolutions. A preprocessed sample spectrum is depicted in Figure 14.

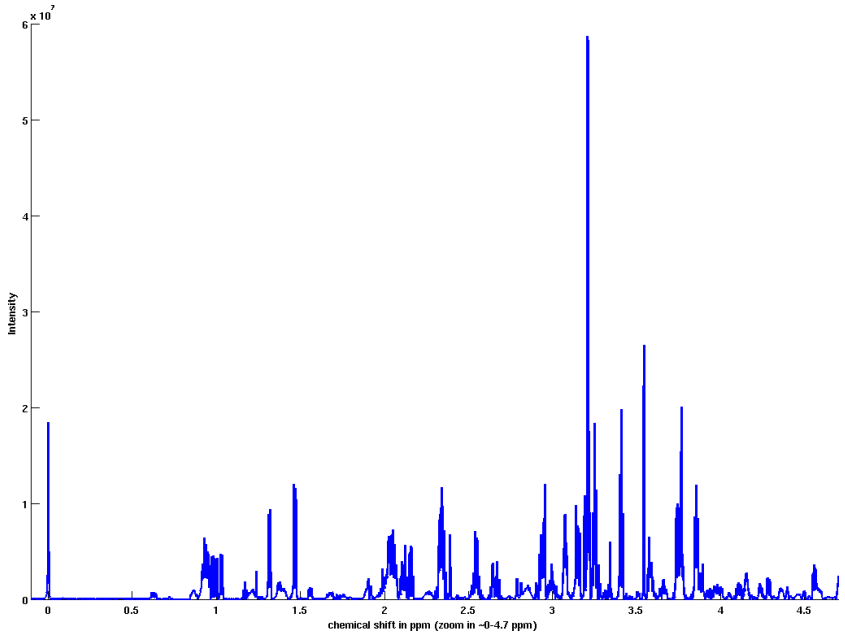


Figure 14: Exemplary preprocessed 1-H NMR spectrum at 700.153 MHz from a typical stem cell extract. The spectrum has been phased and baseline corrected, further the water peak has been removed and a global shift correction has been applied with respect to a provided standard (DSS).

1-H NMR spectra consist of a large amount of relevant signals, such as metabolites which are represented by in general multiple small peaks on top of a wide underlying complex baseline. The NMR signal $s(t)$ can can be roughly approximated as a super composition of Lorentzians [5] but also Gaussian functions or a mixture thereof are common. A setting using

Lorentzians is shown in Eq. 1

$$s(t) = \sum_{j}^{J} A_j e^{i(w_j t+\phi_j)-t/T_{2_j}^*} \quad (1)$$

$$S(w) = \sum_{j}^{J} e^{i\phi_j}(a_j(w)+d_j(w)) \qquad (FFT) \quad (2)$$

with $a_j(w)$ as an absorption signal and $d_j(w)$ as an dispersive signal. This setting however, is very idealized and in practical measurements the line shape of the peaks is much more complex and inhomogeneous due to measurement instabilities or different kinds of coupling effects within the analyzed sample. This generates multiple challenges in the analysis because almost all relevant signals in the NMR measurement showing strong overlapping components. Without an appropriate model of the signal structure a deconvolution is extremely complicated. This is especially true for signal components at low concentrations which maybe easily overlooked otherwise. In the Targeted Profiling (TP) approach [11] such an ideal situation is assumed and it is further assumed that the number of candidate signatures in the mixture $s(t)$ is small and restricted to a specific subset of known metabolites. For such a set of known metabolites (targets) the peak sequence of a plain measurement, e.g. the metabolite Alanine (Ala), maybe known due to theoretical analysis steps incorporating knowledge about the chemical structure of the target and the measurement process. For example the target signal Alanine $f(t)$ can be described as

$$f(t) = \sum_{j}^{G} g_j(t)$$

with $g_j(t)$ as a peak pattern (e.g. a quartet) with appropriate settings for $g_j(t)$ as pointed out later on. An alternative compact description of Alanine is given by

$$\{[1.46^3 ppm, 3.76^1 ppm], [7.234, 7.234, 7.234, -14.366, -14.366, -14.366]\}$$

. Thereby the first part of the pair defines the chemical shifts of the spin descriptors, with 4 spins in this case, whereas the second pair defines the couplings between the pairs in a natural ordering.

It is also possible to generate real measurements of the target and to derive the shape from these experiments. The known peak sequence information (signature) for the metabolites can now be used to analyze the signal with respect to these signatures employing e.g. a partial least squares fit on an appropriate design matrix consisting of the signature information to be expected in $s(t)$. While this approach is quite promising, fast and efficient as pointed out in [11] it suffers from multiple underestimated problems. The main problem comes with the target itself. In TP it is assumed that the signature of the target is perfectly known and can be observed in the signal. This however is not true in general. Due to variations in the measurement, e.g. temperature fluctuations, the positions of the sub patterns in a target (groups of peaks) may shift in a non-linear manner.

Further for the fitting of the targets against the signal a specific line-shape has to be chosen which in general is a Lorentzian or a Gaussian, this however is also a strong assumption which leads to further problems especially for strongly overlapping signals as depicted in Figure 15.

Therefore the phased and baseline corrected signal is better approximated by Eq. 3.

$$s(t) = (\sum_{j}^{J} \alpha_j f_j(t-o)) + \epsilon \tag{3}$$

$$f_j(t) = \sum_{i}^{G} g_i(t-\Delta_i) \tag{4}$$

$$g_i(t) = \sum_{k} \Theta_k(t) \otimes \wp(t) \tag{5}$$

$$\wp = \text{e.g.}\exp(\ldots) \quad \text{line shape} \tag{6}$$

Thereby o can be considered as a global shift which can be compensated by a reference shift correction and ϵ is noise. The target f_j can be approximated as a super composition of its component functions (the contributions of the individual spin-systems). Thereby for each spin-system a small local shift $-\gamma \leq \Delta_i \leq +\gamma$ within a range of typically $|\gamma| \leq 0.005$ ppm can be expected. The components $g_i(t)$ can be considered as line spectra with non vanishing entries for $\Theta_k(t)$ only at one peak position caused by the spin-system due to interactions of carbon-bounded protons in the chemical structure. Subse-

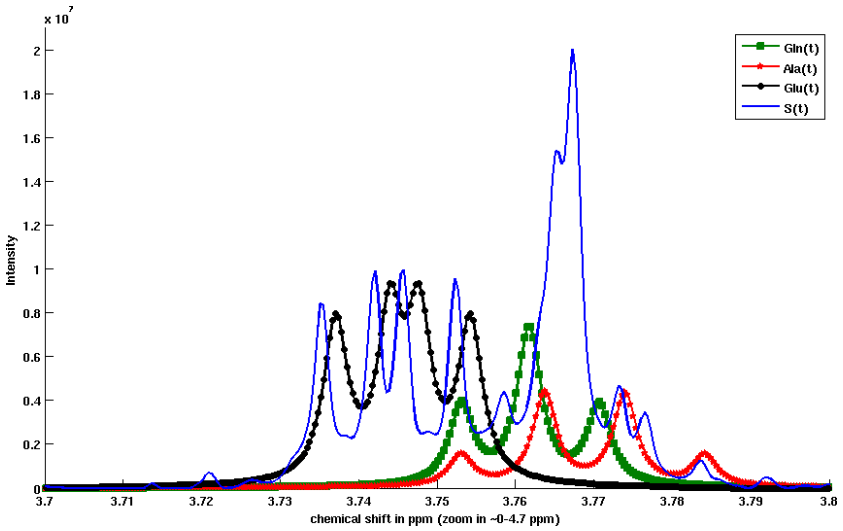


Figure 15: Example of an overlapping effect within a preprocessed 1-H NMR spectrum containing multiple metabolite signatures. It can be clearly observed, that the assumption of the Lorentzian fails in parts to provide a sufficient approximation. This can lead to wrong estimates of target heights and hence wrong concentration estimates.

quently this line spectra is folded $\otimes$ by a line shape function.

In NMR the position of the sub patterns or peaks are known as chemical shifts. The estimates of these shift positions needs to be as accurate as possible and are the main error-source in the TP approach. Further an appropriate peak shape estimate is the key to get an appropriate subtraction of signal components from $s(t)$ and to reveal potentially hidden components. In an initial step in this way our approach takes the shape of the reference signal (CSI - DSS) as a template. This shape is used to estimate e.g. the expected peak width present in the signal and to fit identified targets against $s(t)$.

To overcome the shift problem we estimate values for the disturbances Δ shown in Eq. 3 and present an initial solution to optimize the sub pattern positions in potential targets using a grid search strategy. This approach leads to in general improved position estimates for the *true* chemical shifts of the sub-patterns of potential targets and hence to more accurate identification

and quantification estimates as shown later on.

3 Extended Targeted Profiling

As priorly pointed out Targeted Profiling identifies signatures in NMR mixtures by employing known database references. In the extended Targeted Profiling approach (ETP) we modify this concept such that the spin-system parameters of the targets are optimized with respect to the measurement at hand. Each target description T (corresponding to a signal $f_j(t)$) is characterized by a set of spin-system descriptors $T_d \in S$. S describes the theoretical aspects of the spin system of T and can be used in combination with a model of the measurement system (NMR system) to simulate the spectrum for T. For real measurements of T (e.g. a measurement of alanine) variations of the observed spectrum with respect to the simulation can be observed. This is caused by different effects of the measurement process e.g. inconsistent temperature conditions during the measurement. A spectrum representation of T can be divided into multiple parts, one for each spin-system descriptor T_d, which we will call a peak group (g) subsequently. A peak group may consist of multiple or a single peak and is potentially overlapping e.g due to the measurement resolutions. For each group a potential (limited) shifting error Δ_i can be expected as pointed out priorly. These shift errors are in parts compensated by the following two steps. First an alternative representation of the spectrum by means of a peak representation (line spectra) is generated. Subsequently these observed peak positions are coupled with respect to the known peak position in the simulation such that roughly a assignment of these peaks to a target T and a group g in T is obtained. From these assignments the observed shifts can be calculated and the simulation parameters can be adapted accordingly. Subsequently we detail these two steps. Further a direct analytical approach is briefly described.

3.1 Line representation of a NMR spectrum

NMR spectra can be described by means of a set of overlapping peaks. To generate such a list of peaks, an appropriate model of the peak shape is nec-

essary. In general the peak shape is assumed to be gaussian such that a single peak can be represented by the following equation $\wp(t) = \exp(-(\frac{t-\mu}{\sigma})^2/2)$ with μ as a center position and σ as the line width. Also a Lorenzian peak shape is commonly used as provided before. A further implicit assumption is that the peak shape is symmetric and that the model is sufficient, e.g. is no super composition of gaussians or Lorentzians. In real measurements theses assumptions are only partially filled and a more complex peak shape is observed. This makes the peak picking rather complicated and so far different heuristic approaches have been proposed [5, 2]. Here we focus on a simple parametric hill-climbing approach. We further assume that for each measurement a known reference signal (CSI) is available, in our case this is the DSS signal[1]. This signal has a known position of 0 ppm, which is used to compensate the global shift offset of the spectrum. At the expected DSS position we look (within a window of $0.05ppm$ for a maximum. From this position we go down (to lower intensities) on the left and the right flank of the peak as long as the signal is monotone decreasing. At a predefined maximal width the peak is truncated. For this peak its center position is calculated and the peak width at half maximum (PWHM). The PWHM is used as a rough estimate of the peak width. Due to effects such as imperfect phasing, shimming or baseline correction a direct inverse deconvolution of $s(t)$ with the CSI reference is in general not possible. Instead we employ a hill-climbing algorithm and look for local maxima in the whole signal which are above a predefined threshold (expected noise level), whose flanks are sufficiently steep and for which the obtained peak has a sufficient width. By application of this algorithm we obtain a list of peaks in a spectrum. This list is subtracted from $s(t)$ and the algorithm is repeated until no further peaks are detected. Using this approach also peaks in an overlap can be detected in parts (but not in each case). As an alternative strategy the approach in [5] can be used with an underlying Lorentzian support. The list of peaks is subsequently denoted as $\mathcal{P}$.

[1]The DSS signal consists of more or less a single peak with no or almost no overlap to other peaks. Alternative choices for the CSI such as TSP or ETH are possible as well.

3.2 Peak assignment and shift estimation

In the TP approach only a limited number of spin-system signatures is analyzed with respect to the measurement $s(t)$. In a first step the peak list $\mathcal{P}$ can be filtered such that only those peak positions remain in $\mathcal{P}$ which are part of the peak lists of the targets using a rough shift tolerance of e.g. $0.05ppm$. Now an assignment matrix $M = n \times m$ is generated with n as the number of peaks over all target peaks and m as the number of peaks in $\mathcal{P}$. Thereby multiple assignments are possible and the shift-error of the peak with respect to the expected peak position is stored. Further only such assignments take place which are within a predefined tolerance $0.01ppm$. After this step a voting scheme is applied to M such that a maximal coverage of the target peaks with a minimal error with respect to the shifts is obtained. Hereby its also ensured that a shift applies only to a group but not to the single peaks within a group, because its expected that only the whole group is shifted but the distance between two peaks within g is rather stable. Subsequently one obtains shifts ≥ 0 for each target T and each group g within a target. The line spectra representation of these target peak lists are now optimized with respect to these peak shifts. The optimized target simulations and peak descriptions can now be used in the fitting approach.

3.3 Direct shift analysis by deconvolution

As an alternative to the peak picking approach a more direct solution can be tried. In this case we analyze the regions of the individual groups g of the target by means of a deconvolution problem. Lets assume that the signal range of $s(t)$ is restricted to the range of some overlapping groups g within a tolerance and given as $s'(t)$. Let us further assume that the subset of targets with signature parts (group entries) in the range of $s'(t)$ are T'. Then $s'(t)$ can be considered to be a super-composition of multiple local signatures around the position of g such that

$$s'(t) = \sum_{i}^{|T'|} F(T'_i)$$

with F being a transfer function, generating the spectrum $f(t)$ representation of a local part of a target T. Deconvolution techniques as described in [6] are prominent techniques to separate different overlapping sources. In case of NMR data the positive variants of such deconvolution approaches are preferable in contrast to e.g. ICA [7]. If the number of sources is known and a sufficient amount of measurements is available the signal can be deconvolved into its single sources e.g. using the approach presented in [6]. The obtained source signals can be considered as unmixed contributions to the observed signal. However in practical applications the sources may be still imperfect and noisy such that subsequent processing steps are necessary and the obtain deconvolution model is not directly applicable.

4 Experiments and Results

We analyzed our approach using pure measurements of metabolites and on real life data. Details about the data are shown in Table 1. We also applied the approach on an experimental study of stem cell extracts. For this study stem cells have been cultivated on growing media with different levels of osmolarity and the obtained cell extracts have been analyzed by NMR. For these data a rough assumption on the potential targets is available, which is a necessary condition for (extended) Targeted Profiling, but the ground truth about the identification and concentration of these metabolites is not available. The results of the experiments are listed in the Table 1.

In Figure 16 a reconstruction of a signal part is shown with respect to the original signal to illustrate the effect of the shift correction.

In Table 2 concentrations of measured metabolites are shown, once calculated manually (M) by an expert and again using the ETP approach in an automatic manner (A). The data are obtained from cell extracts, whereby the cells have been cultivated under three different levels of osmolarity (28,32,36 osmol).

Test dataset	Mass	EXP	CTP	ETP
Serine	0.50	0.55	0.38	0.45
Proline	0.31	0.30	0.24	0.09
Malate	0.27	0.27	n.id.	0.25
Alanine	0.38	0.39	0.30	0.33
Glycine	0.41	0.40	0.29	0.41
Threonine	0.28	0.26	0.17	0.28
Mix1-Valine	0.14	0.11	0.04	0.12
Mix1-Inositol	0.14	0.15	0.09	0.08
Mix1-Threonine	0.11	0.12	0.08	0.12
Mix1-Glycine	0.12	0.12	0.10	0.14

Table 1: Comparison of an analysis of pure and mixed metabolites with respect to the true mass, the expert (EXP) analysis, the results provided by CTP and with the new ETP approach. All concentrations are given in μmol.

5 Conclusion

We presented a semi-automatic approach, called, *Extended Targeted Profiling*, for the identification and quantification of metabolites in NMR measurements. Initial results are already quite promising and it could be shown that the new approach is beneficial with respect to traditional techniques and can simplify the metabolite profiling task. Beside of the good agreement of the automatic identification with respect to the expert analysis some challenges remain. For very high overlapping signals our approach is still not accurate enough and manual interactions in the identification task are necessary. In such cases deconvolution approaches have been analyzed briefly and may serve as improved identification techniques in future experiments.

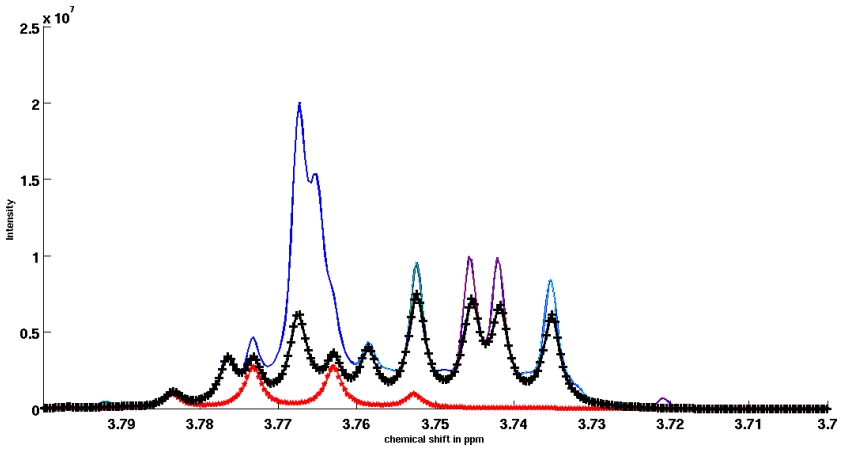


Figure 16: Spectrum with fitted target signals after the extended Targeted Profiling approach. One clearly observes the improved reconstruction (black line, $+$ marker) and the better fit of the individual targets with respect to the function $s(t)$. For Alanine (red line, $\star$ marker) which is completely covered by other signal parts and a hidden underground signal the shift also improved the fit. For Glutamate (causing the quartet on the right) the fit is now almost perfect in comparison to the unshifted fit. Also for Glutamine (signals on the left, $+$ marker) shifting provided improved results. Most of the remaining residual of the signal is now caused due to the (herein) used Lorentzian line shape and one omitted target (large center peak).

Metabolite	28 M	32 M	36 M	28 A	32 A	36 A
Ser	12.66	27.28	40.03	5.68	22.73	24.62
Glu	163.71	289.80	356.12	87.12	159.10	706.46
Pro	23.80	53.29	94.01	37.88	87.12	115.53
Lac	27.97	68.37	89.43	34.09	87.12	107.96

Table 2: Comparison of some results obtained for a real measurement on cell extracts under different osmolarity conditions. All concentrations are given in μmol obtained as an average over three replicates. DSS is chosen as the reference with $18.75\mu\ mol$. While the exact values differ slightly the trend observed by the expert for the different osmolarity levels can be observed for the automatic analysis too.

References

[1] S. Böcker, Matthias C. Letze, Zsuzsanna Liptak, and Anton Pervukhin. Sirius: decomposing isotope patterns for metabolite identification†. *Bioinformatics*, 25(2):218–224, 2009.

[2] G. Brelstaff, Manuele Bicego, Nicola Culeddu, and Matilde Chessa. Bag of peaks: interpretation of nmr spectrometry. *Bioinformatics*, 25(2):258–264, 2009.

[3] D. Chang, C. D. Banack, and S. L. Shah. Robust baseline correction algorithm for signal dense nmr spectra. *Journal of Magnetic Resonance*, 187:288––292, 2007.

[4] M. Cross, R. Alt, and D. Niederwieser. The case for a metabolic stem cell niche. *Cells Tissue Organs*, 188(1-2):150–159, 2008.

[5] H.-W. Koh, J. Lambert, S. Maddula, R. Hergenröder, and L. Hildbrand. Feature selection by lorentzian peak reconstruction for 1-h nmr post processing. In *Proc. of CBMS 2008*, pages 608–613. IEEE Press, 2008.

[6] K. Labusch, E. Barth, and T. Martinetz. Learning data representations with sparse coding neural gas. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks ESANN*, page in press. d-side publications, 2008.

[7] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, New York, 2001.

[8] Pedro Mendes. Metabolomics and the challenges ahead. *Briefings in Bioinformatics*, 7(2):172, 2006.

[9] S. Moco, R. J. Bino, Ric C.H. De Vos, and J. Vervoort. Metabolomics technologies and metabolite identification. *Trends in Analytical Chemistry*, 26(9):855––866, 2007.

[10] F.-M. Schleif. Preprocessing of nuclear magnetic resonance spectrometry data. *Machine Learning Reports*, 1(MLR-01-2007), 2007. ISSN:1865-3960 http://www.uni-leipzig.de/compint/mlr/mlr_01_2007.pdf.

[11] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky. Targeted profiling: Quantitative analysis of 1h nmr metabolomics data. *Analytical Chemistry*, 78:4430–4442, 2006.

[12] Y. Xi and D. M Rocke. Baseline correction for nmr spectroscopic metabolomics data analysis. *BMC Bioinformatics*, 9:324–333, 2008.

[13] J. Xia, T. C. Bjorndahl abd P. Tang, and David S Wishart. Metabominer – semi-automated identification of metabolites from 2d nmr spectra of complex biofluids. *BMC Bioinformatics*, 9:507–522, 2008.

[14] Q. Zhao, R. Stoyanova, S. Du, P. Sajda, and T. R. Brown. Hires—a tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics*, 22(20):2562–2564, 2006.

Deconvolution and identification of mass spectra from mixed and pure colonies of bacteria

Stephan Simmuteit*[1,2]*, Jessica Simmuteit*[3]*,

***Frank-Michael Schleif*[2]*, Thomas Villmann*[4]

1 Introduction

The identification of bacteria is a very basic task in clinical and scientific environments. Biochemical methods like gram staining are commonly used, but time consuming and in some cases not specific enough. Also colonies of different bacteria are not separable by this technique.

Mass spectrometry (MS) provides unique molecular fingerprints of the peptide composition of bacteria [1]. This technology is fast, cheap and reproducible, but one has to deal with high-dimensional spectra, containing noise and distortions of the measurement process. A typical MS measurement of the bacterium *Vibrio Harveyii* is shown in Figure 17. Several pre-processing steps are done, including denoising, baseline subtraction and peak picking. The pre-processing results in a line spectrum, which contains the peak intensities at their respective masses. Here we present a method for the process-

[1]E-mail: simmuteit@googlemail.com
[2]Leipzig University AGCI, Semmelweisstraße 10, 04103 Leipzig
[3]Technical University Clausthal, 38678 Clausthal
[4]University of Applied Sciences Mittweida, 09648 Mittweida

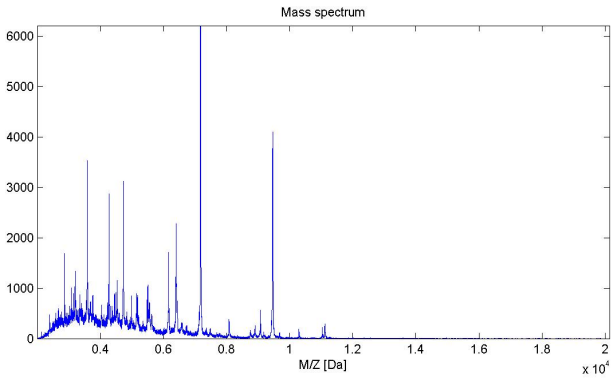


Figure 17: Mass spectrum of cultivated *Vibrio Harveyii*. The x-axis shows the m/z-value in Dalton (Da), the y-axis is an unit less intensity.

ing of mass spectrometric data taken from bacteria samples and derive an algorithmic approach to represent such data by means of an identification model. Aspects of mixed spectra, caused by mixed bacteria are considered as well as appropriate representation techniques employing the bacteria taxonomy. It is also shown how these models can be used to identify bio-patterns characterizing individual clusters of data.

Known identification approaches are based on the direct comparison of given reference spectra and the unknown spectrum [5]. The determination of relevant masses for clustering of a given set of spectra is approached in this contribution as well as the segregation of spectra, which are measured from bacterial mixtures. The first topic is accomplished by a hierarchical variant of the Self-Organizing Map (SOM) [2] called *Evolving Tree* (ET)[7] in combination with the Oja-learning rule [6], providing a local hierarchical principal component analysis. The approach to the second topic is based upon *Sparse Coding Neural Gas* (SCNG) and *Orthogonal Matching Pursuit* (OMP)[4].

2 Evolving Trees and hierarchical PCA

Figure 18 shows spectra from *Listeria Ivanovii Ivanovii* and *Listeria Ivanovii Londonsiensis*. They are almost identical except for intensity variations and

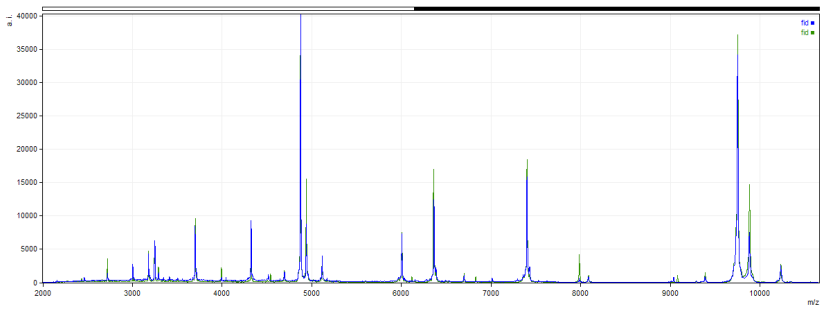


Figure 18: Mass spectra of *Listeria Ivanovii Ivanovii* and *Listeria Ivanovii Londonsiensis*

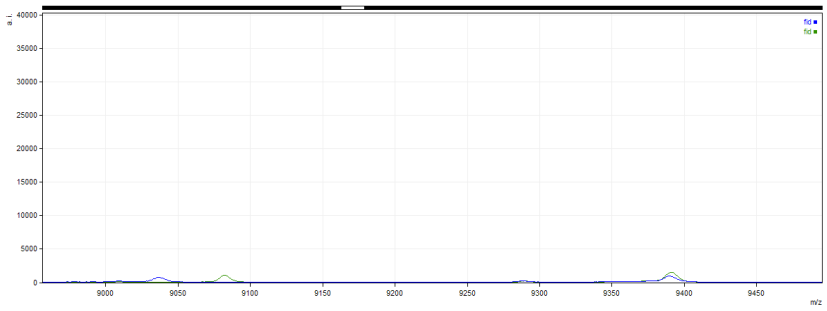


Figure 19: Mass spectra of *Listeria Ivanovii Ivanovii* and *Listeria Ivanovii Londonsiensis* with magnification of the mass area around 9037 Da.

noise. The only real difference can be found at the mass position 9037Da. This is visible in Figure 19.

Respecting the taxonomic nature of bacteria, the representation of MS data should consequently be tree structured. Simple decision trees do not fulfill the requirements regarding data shape and density, but ETs are capable to provide an adequate model, because they are derived from Self-Organizing maps known to be efficient in presenting high-dimensional data with complex data shapes and distributions[2]. Additionally the ET is combined with the Oja-learning rule to achieve a more compact representation of the high-dimensional data, highlighting the most relevant dimensions in a PCA like manner. The introduced Oja-ET provides an inherent principal component

analysis of the represented data, i.e. every prototype vector becomes the first eigenvector of its cluster.

An ET is a growing SOM with a tree shaped neighborhood. It is a mapping of N prototypes $i \in A$ with weight $w_i \in \mathbb{R}^d \supseteq M$ and A as the predefined SOM-grid onto a best matching unit (BMU) i^*. M is the space of the training data. The prototypes are mapped onto a weight vector, which are element of M and d as the dimensionality of the data. These two mappings are explained in equation (1) and (2).

$$M \rightarrow A : \mathbf{x} \in M \mapsto i^*(\mathbf{x}) \in A \tag{1}$$

$$A \rightarrow M : i \in A \mapsto \mathbf{w_i} \in M \tag{2}$$

The Oja-update is shown in equation (3).

$$\Delta \mathbf{w_i} = \alpha h_{i^*,i} O \left(\mathbf{x} - O\mathbf{w_i}\right) \quad \forall i \in A \tag{3}$$

with

$$O =< \mathbf{x}, \mathbf{w_{i^*}} > \tag{4}$$

α is the learning rate which decreases logarithmically in the learning iterations. $h_{i^*,i}$ is the neighborhood function, which also decreases while learning and depends on the distance between i and i^* in the tree structure. The ET learning starts with an initial small tree and if one prototype reaches a predefined threshold of being the BMU, this prototype adds an amount of child nodes and remains itself fixed, i.e. not updated any further. Details on the complete algorithm can be found in [8]. The winner determination by means of Oja-rule is shown in equation (5).

$$i^*(\mathbf{x}) = \arg\max_{i \in A} \langle \mathbf{x}, \mathbf{w_i} \rangle \tag{5}$$

Figure 20 shows an example of a tree with different Listeria species. It is possible to analyze the loadings , i.e. the contributions of single dimensions to the first principal component, of the Oja-ET to get insight into the clustering decision. In Table 1 the analysis for the root node of an Oja-ET with 231 measurements from 7 different *Listeria* species is shown. The first entry reflects

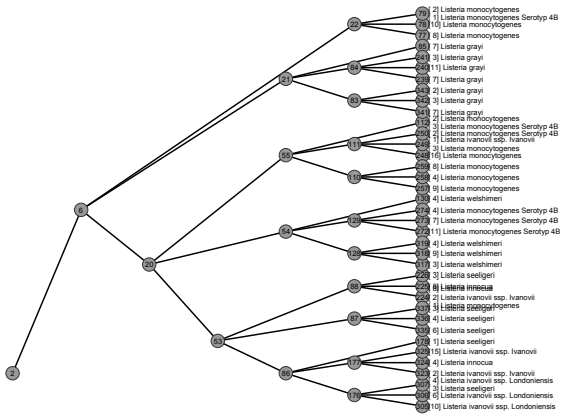


Figure 20: Evolving Tree visualization of different Listeria species.

the importance of the mass position 9751.1Da. The peak at this mass separates *Listeria Grayi* from all other *Listeria* species. The above mentioned peak, which separates *Listeria Ivanovii ivanovii* and *Listeria Ivanovii londonsiensis* can be found on rank 14 in this table, although the separation in the tree takes place later. The table ranks the first from the total of 1181 occurring peaks in this data set and the last-ranked dimension.

3 Source Separation with Sparse Coding Neural Gas

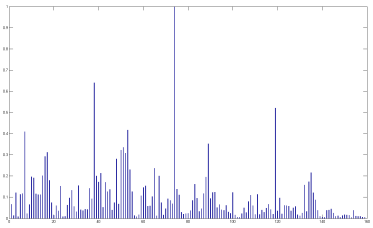


Figure 21: Mixed signals from *Enterobacter Cloacae* and *Acetobacter baumanii*

If a mass spectrum is a measurement of $M \in \mathbb{N}$ mixed bacterial cultures, the presented identification approaches are not able to compute reliable re-

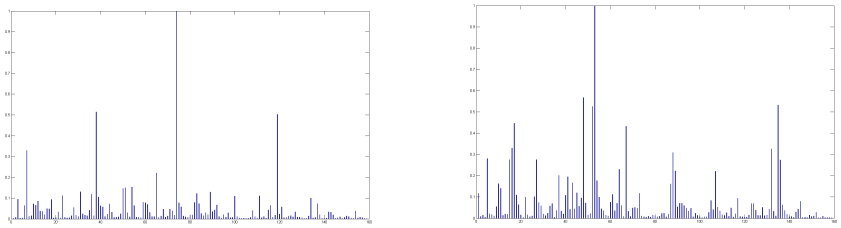


Figure 22: Pure culture signals from *Enterobacter Cloacae* and *Acetobacter baumanii*

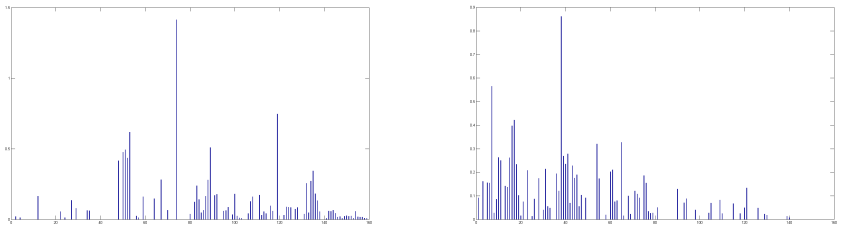


Figure 23: Calculated basis signals from *Enterobacter Cloacae* and *Acetobacter baumanii*

Rank	Contribution	Mass	Rank	Contribution	Mass
1	0.6889	9751.1	8	0.0905	9883.4
2	0.5660	4876.1	9	0.0890	6387.1
3	0.2309	7402.2	10	0.0838	4942.3
4	0.2055	6362.7	11	0.0743	9714.4
5	0.1947	4323.3	12	0.0693	3249.8
6	0.1184	6006.8	13	0.0466	3181.4
7	0.1045	3700.9	**14**	**0.0424**	**9036.9**
			..	..	..
			1181	$4.2 \cdot 10^{-12}$	5418.5

Table 1: Analysis of the loading of the root node of an Oja-ET with Listeria spectra. The most important mass is 9751.1Da, separating *Listeria Grayi* from the other Listeria species.

sults. An obvious idea is the deconvolution of the mixed signal. We assume, that the mixed spectra turn out to be a linear combination of some basis spectra representing pure bacteria cultures. Hence one can try to represent a spectrum as a sum of basis functions. The SCNG algorithm [4] assumes M sparse sources $S = (s_1, \ldots, s_M)^T = (a_1, \ldots, a_L)$ and N observations. The observations $\mathbf{X} = (\mathbf{x_1} \ldots \mathbf{x_L}), \mathbf{x_j} \in \mathbf{V} \subseteq \mathbb{R}^n$ should be represented by (6).

$$\mathbf{x_j} = \mathbf{C}\mathbf{a_j} + \epsilon_j \tag{6}$$

with $\mathbf{C} = (\mathbf{c_1}, \ldots, \mathbf{c_M}), \mathbf{c_j} \in \mathbb{R}^n$. The vector $\mathbf{a_j}$ contains the contribution of the source $\mathbf{s_i}$ to the mixture $\mathbf{x_j}$. ϵ_j is a noise term. Together with the OMP-Algorithm it is now possible the reconstruct the basis functions from the mixtures and the mixing matrix determined by SCNG[3].

Experiments are done with artificial data as well as with real measurements of mixtures of *Enterobacter Cloacae* and *Acetobacter baumanii*. Initial results show an observable separation of the different species. An example is shown in Figure 21 showing a mixture of the above mentioned bacteria. Pure cultures measurements are shown in Figure 22. Calculated basis spectra are shown in Figure 23. The first results of the approach show promise and will be extended in future work.

4 Conclusion

A method for unsupervised hierarchical clustering of mass spectrometry data from bacteria has been present. It provides an inherent local hierarchical principal component analysis, which is used to identify relevant masses for the clustering process. We also obtain an interpretable representation of the spectra, which reflects the taxonomical nature of the bacteria species. The model allows a retrieval of unknown data with logarithmic costs. The problem of identifying different bacteria in one spectrum was approached by Sparse Coding Neural Gas, which breaks a spectrum into a set of basis spectra, providing the spectra, which would have been measured, if the spectra would have been measured separately. Results show a very good detection of important peaks for bacteria species separation, as well as a visual impression of deconvoluted pure bacteria spectra.

References

[1] S. B. Barbuddhe, T. Maier, G. Schwarz, M. Kostrzewa, H. Hof, E. Domann, T. Chakraborty, and T. Hain. Rapid identification and typing of listeria species by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Applied and Environmental Microbiology*, 74(17):5402–5407, 2008.

[2] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (2nd Ed. 1997).

[3] K. Labusch, E. Barth, and T. Martinetz. Learning data representations with sparse coding neural gas. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks ESANN*, pages 233–238. d-side publications, 2008.

[4] Kai Labush, Erhardt Barth, and Thomas Martinetz. Sparse coding neural gas for the separation of noisy overcomplete sources. In *Artificial Neural Networks - ICANN 2008, PT I*, volume 5163 of *Lecture Notes in Computer Science*, pages 788–797, 2008.

[5] T. Maier and M. Kostrzewa. Fast and reliable MALDI-TOF MS-based microorganism identification. *Chemistry Today*, 25:68–71, 2007.

[6] Erkki Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.

[7] Jussi Pakkanen, Jukka Iivarinen, and Erkki Oja. The evolving tree—a novel self-organizing network for data analysis. *Neural Process. Lett.*, 20(3):199–211, 2004.

[8] Stephan Simmuteit. Effizientes Retrieval aus Massenspektrometriedatenbanken, Diplomarbeit, Technische Universität Clausthal. February 2008.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

Relevance learning for generative topographic maps

Andrej Gisbrecht[1,2]

1 Introduction

Rapidly developing technology such as improved sensor technology and high resolution of imaging techniques have turned the analysis of very high dimensional data a central issue of data mining. As a consequence, many techniques which help humans to inspect these data have emerged in the past years, see e.g [4] for an overview. As an example, mass spectrometric instruments become more and more sensitive leading to a dimensionality of 10.000 and more of the raw data. One example is shown in Fig. 24. This displays the outcome of a survey on influenza vaccination made by East Virginia Medical School, where the original data have been preprocessed using peak detection techniques. The goal of this study was the investigation of the response to vaccination based on different factors. In particular, the dependency of the response on the age of the patient is a critical issue, since, probably, the immune system of older people could already be too weak to respond. For the study, the outcome of the vaccine (responder/non-responder) was known for all patients. Obviously, these data yield high dimensional time series which are labeled according to the known outcome.

The motivation of our work is to provide machine learning tools which

[1]E-mail: andrej.gisbrecht@tu-clausthal.de
[2]Department of Informatics, Clausthal University of Technology, Clausthal, Germany

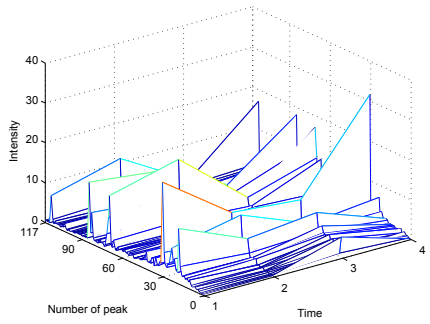


Figure 24: Dataset displaying preprocessed mass spectrometric time series taken in the course of a study on the response to vaccines.

allow to inspect high-dimensional data when auxiliary information such as class labels is available. Thereby, one goal is to obtain a scheme which weights input factors according to their relevance for the output label such that only the relevant information is displayed in this setting. Further, an interpretation of the data and the relevance of the single input features for this map should be possible.

Our work is based on the Generative Topographic Mapping by [2][5] which provides a well-founded stochastic model for data visualization. This model is extended by relevance learning based on Hebbian ideas as proposed e.g. in the approach [3] for supervised prototype-based classification schemes.

2 Generative Topographic Mapping

The Generative Topographic Mapping (GTM) has been introduced by Bishop et al. [2]. There exists an enhancement of this algorithm [1] which can deal with the time series character of the data. However, for the moment, we focus on classical GTM. GTM is similar to the Self-Organising Map algorithm with the difference, that it is based on a probabilistic model. The algorithm tries to explain the distribution of given high dimensional data points by a small number of latent points in a low dimensional mapping space, as displayed in Fig. 25.

Assume N data points in the real D-dimensional data space are given. The

distribution of these points should be explained by K points in an L-dimensional latent space, which are ordered on a grid. The GTM defines a parametric mapping $y(\mathbf{x}, \mathbf{W})$ from the latent space to the data space. In this way the latent points are embedded in the L-dimensional manifold in the data space. The data points, which lie in the Gaussian bell around an embedded latent point, are probably induced by this latent point. We define the probability, that the data point $\mathbf{t}$ was generated by the latent point $\mathbf{x}$ as

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2}\sum_{d}^{D}(t_d - y_d(\mathbf{x}, \mathbf{W}))^2\right),$$

with β^{-1} defining the bandwidth of the Gaussian curve.

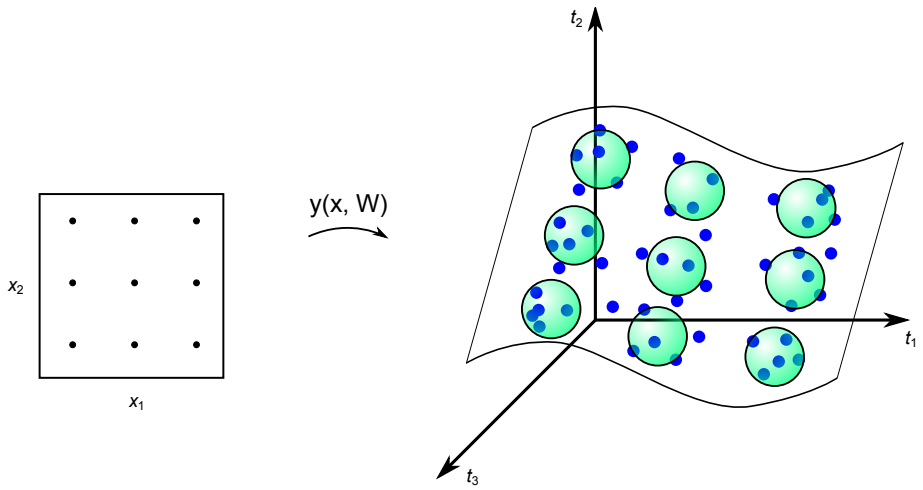


Figure 25: Mapping from the latent to the data space as given in the GTM model

The probability that data point $\mathbf{t}$ was generated by the model is calculated as the sum over all latent points

$$p(\mathbf{t}|\mathbf{W}, \beta) = \frac{1}{K}\sum_{k}^{K} p(\mathbf{t}|\mathbf{x}^k, \mathbf{W}, \beta).$$

The likelihood function for the model is given by the product

$$F = \prod_{n}^{N} \left(\frac{1}{K} \sum_{k}^{K} p(\mathbf{t}^n | \mathbf{x}^k, \mathbf{W}, \beta) \right)$$

assuming independence of the samples. During training, the likelihood function is maximised using an expectation-maximization (EM) algorithm. For each latent point, the EM algorithm calculates the probability that the point was generated by this latent point. The resulting responsibilities r^{kn} are

$$r^{kn} = \frac{p((\mathbf{t}^n | \mathbf{x}^k, \mathbf{W}, \beta))}{\sum_{k'} p(\mathbf{t}^n | \mathbf{x}^{k'}, \mathbf{W}, \beta)}.$$

The parameters of the mapping y are changed in such a way, that the latent points are moved to the data points for which they possess high responsibility.

3 Supervised Relevance GTM

Our aim is to integrate supervised label information into GTM and to adapt the underlying euclidean metric according to this auxiliary information. For this purpose, we use ideas as introduced in [3].

For supervised relevance GTM (SRGTM), the Euclidean distance is enlarged by relevance factors for all dimensions

$$|\mathbf{a} - \mathbf{b}|_\lambda^2 = \sum_{d}^{D} \lambda_d^2 (a_d - b_d)^2$$

The idea behind this setting is that, for a high dimensional problem, some dimensions are probably less relevant for the classification or just noise. An appropriate scaling of those dimensions would help to suppress noise in the data which does not affect the classification and which should not be displayed by GTM.

The relevance terms λ_d should be adjusted automatically during training. For this purpose, we attach a labeling to latent points of GTM with

$$L^k = \arg\max_{l} \left(\sum_{n} r^{kn} | l^n = l \right)$$

where L^k is the label of the k-th latent point, l^n is the label of the n-th data point and r^{kn} are the responsibilities. This allows to transfer Hebbian learning to supervised GTM as follows: The update rule for λ for all k, n and d is

$$\lambda_d = \begin{cases} \max\left(0, \lambda_d - \eta \cdot \lambda_d \cdot r^{kn} \left(t_d^n - y_d(\mathbf{x}^k, \mathbf{W})\right)^2\right) & \text{if } L^k = l^n \\ \lambda_d + \eta \cdot \lambda_d \cdot r^{kn} \left(t_d^n - y_d(\mathbf{x}^k, \mathbf{W})\right)^2 & \text{if } L^k \neq l^n \end{cases}$$

After the update, λ is normalised to prevent the degeneration. This update is reasonable, since, in case the labels of a latent point and a data point are the same, the dimensions with big variance are probably not important and their relevance is decreased. If the labels are different, then the importance of the dimensions with small variance is increased only slightly, thus, after normalisation, the relevance of the dimensions which contribute most to the wrong classification is decreased most.

4 Data evaluation

We investigate the behavior of GTM and SRGTM on an artificial data set given in [3]. The first two dimensions of the data are plotted in figure 26. The dimensions 3 to 6 are noisy copies of the first dimension and the dimensions 7 to 10 consist only of noise. The dataset has 3 classes, each with 2 clusters. Red and Blue classes have significant overlap. The accuracy of RLVQ on this dataset is about 0.8.

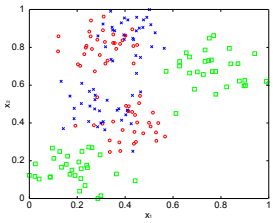
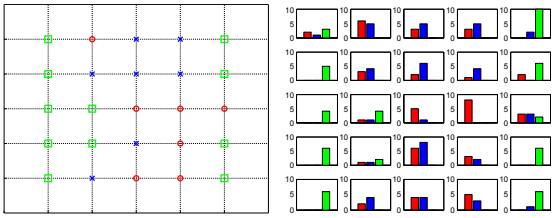


Figure 26: Dataset

On the figure 27 the trained GTM-map is plotted together with the distribution of the classes to the latent points. As we see, the map visualises the

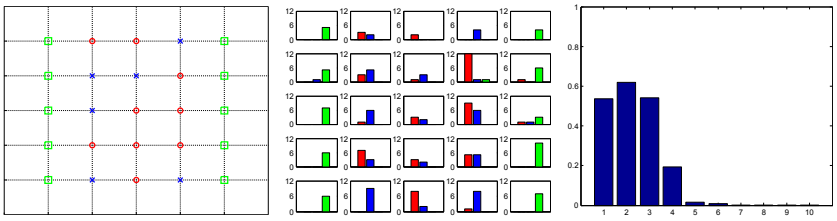
original data. The green class is on the sides and has small overlap with blue and red classes. The red and blue classes are in the middle overlapping each other. The classification accuracy with leave-one-out cross validation is 0.71.



(a) Visual structure (b) Distribution of the classes

Figure 27: Visual structure of the GTM-map

SRGTM achieves an accuracy of 0.82, which is better than GTM and comparable to RLVQ. Obviously, the SRGTM-map, plotted in the figure 28, represents the data set better than the GTM-map. The green class is clearly separated from the red and blue classes. The red and blue classes are again overlapping in the middle, but each latent point is mostly responsible for only one class.



(a) Visual structure (b) Distribution of the classes (c) Relevances of the dimensions

Figure 28: Visual structure of the SRGTM-map

Additionally to the visualisation of the data the SRGTM provides the relevances of the dimensions. As we can see in figure 28(c), the first two dimensions are clearly relevant. The dimensions 3 and 4 are relevant as well, since they are noisy copies of the first dimension. The dimensions 5 and 6

are not relevant, because the added noise is bigger than for dimensions 5 and 6. Dimensions 7 to 10, which contain pure noise, are not relevant at all.

References

[1] C.M. Bishop, G.E. Hinton, I.G.D. Strachan *GTM Through Time.* Proceedings IEE Fifth International Conference on Artificial Neural Networks, Cambridge, U.K., pp.111-116, 1997.

[2] C.M. Bishop, M. Svensen, C.K.I. Williams *GTM: The Generative Topographic Mapping.* Neural Computation 10:215-234, 1998.

[3] B. Hammer, T. Villmann. *Generalized relevance learning vector quantization.* Neural Networks 15, 1059-1068, 2002.

[4] D.A. Keim Information Visualization and Visual Data Mining IEEE Transactions on Visualization and Computer Graphics 8(1):1-8, 2002.

[5] J.F.M. Svensen. *GTM: The Generative Topographic Mapping.* PhD thesis. Aston University, 1998.

Matrix learning and data visualization

Barbara Hammer[1,2]

1 Introduction

The amount of electronic data available today doubles roughly every twenty months. At the same time, the data dimensionality increases tremendously due to highly improved technical processes such as the improved resolution of imaging techniques or the high sensitivity of sensors. As a consequence, automatic methods are needed which allow humans to rapidly scan through these data sets. One possibility is offered by efficient methods for data visualization which allow humans to visually inspect these data sets and to easily detect important structures such as clustering or outliers in the data.

The research topic of dimensionality reduction and data visualization has emerged rapidly in the last years, see e.g. [7, 5] for recent overviews. Since data visualization is an inherently ill-posed problem, a variety of different objectives have been proposed which yield different visualization of a given data set, including, for example, the principle of metric preservation, preservation of the local data topology, or minimization of the information loss in a least squares sense, to name just a few. Two problems can frequently be observed in these settings: methods are often computationally complex and hardly applicable to huge data sets because of superlinear dependency on the number of data; further, complex nonlinear methods provide a mapping of given data rather than an explicit embedding function, such that out

[1]E-mail: hammer@in.tu-clausthal.de
[2]Institute of Computer Science, Clausthal University of Technology, Germany

of sample extensions to new data points are difficult.

Here, we focus on extensions of prototype based approaches to data visualization. By representing the data characteristics in a fixed number of prototypes, these methods lead to algorithms which scale only linearly with the size of the data set. Recent extensions of prototype based methods allow to automatically adapt the underlying metric by means of an adaptive matrix attached to every prototype. This constitutes a key issue for direct data visualization: the local matrices give rise to local linear projections of the data, which can be combined to a global mapping using a standard technique such as charting. This way, an explicit mapping of the data into low dimensions is obtained. We demonstrate this pipeline within an unsupervised matrix learning framework, yielding to unsupervised manifold embedding, as well as a supervised matrix learning environment, which allows to obtain a discriminative nonlinear embedding of the underlying data manifold in low dimensions.

2 Matrix learning

Prototype based methods represent data $\vec{x}_1, \ldots, \vec{x}_m$ in $\mathbb{R}^n$ by prototypes $\vec{w}_1, \ldots, \vec{w}_p$ in the same space. The assignments are given by a metric based winner takes all scheme which assigns the data point $\vec{x}_i$ to its closest prototype $\vec{w}_j$ with $d(\vec{x}_i, \vec{w}_j)$ minimum. Thereby, often, the squared euclidian distance

$$d(\vec{x}_i, \vec{w}_j) = (\vec{x}_i - \vec{w}_j)^t(\vec{x}_i - \vec{w}_j)$$

is used. There exist supervised as well as unsupervised machine learning techniques to determine appropriate prototypes from a given data set, aiming at clustering or classification of the data space, respectively.

Neural gas (NG) constitutes one example for a very robust unsupervised clustering technique [8]. The aim of NG is the minimization of the cost function

$$\sum_{ij} \exp(-\mathrm{rk}_{ij}/\sigma^2) d(\vec{x}_i, \vec{w}_j)$$

where

$$\mathrm{rk}_{ij} = |\{\vec{w}_k \mid d(\vec{x}_i, \vec{w}_k) < d(\vec{x}_i, \vec{w}_j)\}|$$

denotes the rank of prototype $\vec{w}_j$ if arranged according to its distance from data point $\vec{x}_i$ and σ indicates the strength of neighborhood cooperation. Training algorithms can be derived thereof using a stochastic gradient descent (online NG) or a batch approach (batch NG).

A popular supervised prototype-based classification scheme is offered by learning vector quantizers (LVQ). In LVQ, every prototype $\vec{w}_j$ is equipped with a class label $c(\vec{w}_j)$, and a data point $\vec{x}_i$ is mapped to the class of the closest prototype, its winner. Since the classification error cannot easily be optimized directly, a variety of training paradigms have been derived based on alternative objectives. Original LVQ1, LVQ2.1 and LVQIII as proposed by Kohonen [6] rely on heuristics and move prototypes towards/away from a presented data point in Hebbian style depending on whether the classification is correct. Alternatives propose cost functions which are related to the misclassification error. A very robust method is offered by generalized LVQ (GLVQ) [9] which optimizes the cost function

$$\sum_i \Phi\left(\frac{d(\vec{x}_i, \vec{w}^+) - d(\vec{x}_i, \vec{w}^-)}{d(\vec{x}_i, \vec{w}^+) + d(\vec{x}_i, \vec{w}^-)}\right)$$

where $\vec{w}^+$ denotes the closest prototype with the same class label as $\vec{x}_i$ and $\vec{w}^-$ denotes the closest prototype with a different class label than $\vec{x}_i$. Φ denotes a monotonic increasing function such as the logistic function. Since the denominator is negative iff the data point $\vec{x}_i$ is correctly classified, this cost function constitutes a reasonable approximation of the classification error. A stochastic gradient descent yields update formulas which display a large similarity to heuristically based LVQ schemes, while offering a sound mathematical derivative.

Both methods, supervised GLVQ and unsupervised NG rely on the choice of the distance measure. The choice of d as squared euclidian distance induces cluster shapes which stem from isotropic isobars. For real life problems, this is often not the best choice and it turns out particularly crucial if high dimensional and noisy data are dealt with. As a consequence, methods to substitute the euclidian metric by a more general choice and technology to adapt the metric automatically according to the given data at hand have been proposed. For both, GLVQ and NG, automatic matrix learning schemes have been proposed recently [9, 2]. The basic idea is to substitute the squared

euclidean metric by the choice

$$d(\vec{x}_i, \vec{w}_j) = (\vec{x}_i - \vec{w}_j)^t \Lambda_j (\vec{x}_i - \vec{w}_j)$$

with a symmetric positive definite matrix Λ_j attached to prototype $\vec{w}_j$. This generalization allows an adaptation of the local distance measure at prototype $\vec{w}_j$ according to the situation at hand, such that local ellipsoidal isobars determine the clustering or classification, respectively. The resulting decision boundaries are no longer necessarily linear, rather, quadratic surfaces can result.

This more general metric can be directly plugged into the cost function of NG. Enforcing the constraint $\det \Lambda_j = 1$ to prevent divergence to a trivial solution, a batch update scheme can be derived which sets the matrices to a generalized inverse local covariance matrix of the data. Hence, the matrices are automatically symmetric and positive definite, i.e. they define a valid metric. Further, they take the local data statistics and local cluster shape into account since the covariance matrix can be related to local PCA schemes.

For GLVQ, the constraints on the matrix can be implemented by setting $\Lambda_j = \Omega_j \Omega_j^t$ such that a stochastic gradient descent can be derived also for matrix updates. Thereby, degeneration has to be prevented e.g. enforcing $\mathrm{trace}(\Lambda_j) = 1$. Both methods yield promising results when used for clustering or classification, as demonstrated in [9, 2].

3 Manifold charting for data visualization

Manifold charting has been proposed in the approach [3] as a low dimensional embedding technique which yields an explicit map for the embedding. The method consists of two steps. Step one determines local linear projections of the data to low dimensions by centering a local PCA scheme at every data point of the given data set. Step two combines these local projections such that a uniform global nonlinear mapping results. Basically, the local linear projections are glued together by affine transformations of the local pieces such that they coincide as much as possible on the overlaps in a least squares sense. It has been shown in [3] that the coefficients of this map can be computed analytically.

Matrix learning in NG and LVQ provides local linear maps centered around the prototypes by a projection to the most important principal components of the local matrices. Thereby, the most important principal components are the components according to the smallest eigenvalues of the matrix corresponding to the biggest variances of the data for unsupervised schemes such as NG. Conversely, the most important directions are the ones corresponding to the largest eigenvalues for supervised schemes because these explain the major parts of the classification. Thus, from supervised as well as unsupervised prototype based matrix learning, p local projections into low dimensions can be gained, p denoting the number of prototypes. These mappings can be glues together using directly the second step of manifold charting, whereby the responsibilities of the local linear maps are given by Gaussians centered around the prototypes. As a result, a global visualization results in both cases, which focusses on the directions of largest variance in the unsupervised case and the most discriminative directions in the supervised case. Since the method relies on p local maps only, the overall procedure is linear in the number of training points. Further, explicit functions are obtained for the mapping in both cases.

First promising results of these methods have been obtained in a couple of artificial and real life benchmarks, see [4, 1]. Interestingly, the charting technique based on supervised matrix learning offers one of the few nonlinear discriminative data visualization techniques with direct out of sample extensions and only linear effort.

References

[1] B. Arnonkijpanich, B. Hammer, and A. Hasenfuss. *Local matrix learning in clustering and applications for manifold visualization*. Submitted.

[2] B. Arnonkijpanich, B. Hammer, A. Hasenfuss, and C. Lursinsap. Matrix Learning for Topographic Neural Maps. in: Vera Kurkova, Roman Neruda, and Jan Koutnik (eds.), ICANN (1), pp. 572–582, 2008. Springer.

[3] M. Brand. *Charting a manifold*. Tech. Rep. 15, Mitsubishi Electric Research Laboratories (MERL),

URL http://www.merl.com/publications/TR2003-013/, http://www.merl.com/publications/TR2003-013/, 2003.

[4] K. Bunte, B. Hammer, and M. Biehl. *Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data*. Submitted

[5] D.A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[6] T. Kohonen. *Self-organizing Maps*. Springer, 2001.

[7] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality Reduction: A Comparative Review. Submitted, 2009.

[8] T.M. Martinetz, S.G. Berkovich, and K.J. Schulten. Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):-569, 1993.

[9] P. Schneider, M. Biehl, and B. Hammer. Matrix learning in learning vector quantization. *Neural Computation*. To appear.

Topographic mapping techniques for dissimilarity datasets

Alexander Hasenfuss[1,2]***, Barbara Hammer***[1]

1 Introduction

The presence of huge data sets, often several GB or even TB, brings special challenges to standard data clustering and visualization techniques, such as Neural Gas (NG) or the Self-Organizing Map (SOM) [20, 8]. At most a single pass over the data is affordable therefore online adaptation which requires several runs over the data is not applicable. At the same time, alternative fast batch optimization cannot be applied due to memory constraints. In recent years, researchers have worked on so-called single pass clustering algorithms which run in a single or few passes over the data and which require only a priorly fixed amount of allocated memory. Popular methods include heuristics like CURE, STING, and BIRCH [13, 15, 16] and approximations of k-means clustering as proposed in [12, 21]. In addition, dynamic methods such as growing neural gas have been adapted to cope with the scenario of life-long adaptivity, see e.g. [22].

The situation becomes even more complicated when data are non-vectorial and distance-based clustering methods have to be applied, which often feature a quadratic time complexity [6]. Although a variety of methods which can directly work with relational data based on general principles such as

[1]Clausthal University of Technology, Department of Informatics, Clausthal-Zellerfeld, Germany
[2]E-mail: hasenfuss@in.tu-clausthal.de

extensions of the self-organizing map and neural gas have been proposed [19, 3, 14], these methods are not suited for huge data sets. For complex metrics such as alignment of DNA strings or complex kernels for text data, it is infeasable to compute all pairs of the distance matrix and at most a small fraction can effectively be addressed. A common challenge today [10], arising especially in Computational Biology, are huge datasets whose pairwise dissimilarities cannot be hold at once within random-access memory during computation, due to the sheer amount of data.

In the presentation, we introduced a technique based on the Relational NG and Relational SOM approach [14] that is able to handle this situation by a single pass technique based on patches that can be chosen in accordance to the size of the available random-access memory. This results in a linear time and constant memory algorithm for general dissimilarity data which shares the intuitivity and robustness of NG and SOM.

2 The Relational Approach

Relational data do not necessarily originate from an Euclidean vector space, instead only a pairwise dissimilarity measure d_{ij} is given for the underlying datapoints $v_i, v_j \in V$. The only demands made on dissimilarity measures are non-negativity $d_{ij} \geq 0$, reflexivity $d_{ii} = 0$, and symmetry $d_{ij} = d_{ji}$, so they are not necessarily metric by nature.

One way to deal with relational data is Median clustering [3]. This technique restricts prototype locations to given data points, such that distances are well defined in the cost function of NG. Batch optimization can be directly tranferred to this case. However, median clustering has the inherent drawback that only discrete adaptation steps can be performed which can dramatically reduce the representation quality of the clustering.

Relational Neural Gas (RNG) [14] overcomes the problem of discrete adaptation steps by using convex combinations of Euclidean embedded data points as prototypes. For that purpose, we assume that there exists a set of (in general unknown and presumably high dimensional) Euclidean points V such that $d_{ij} = \|v_i - v_j\|$ for all $v_i, v_j \in V$ holds, i.e. we assume there exists an (unknown) isometric embedding into an Euclidean space. The key observation

is based on the fact that, under the assumptions made, the squared distances $\|w_i - v_j\|^2$ between (unknown) embedded data points and optimum prototypes can be expressed merely in terms of known distances d_{ij}. This allows to reformulate the batch optimization schemes in terms of relational data as done in [14].

Note that, if an isometric embedding into Euclidean space exists, this scheme is equivalent to Batch NG and it yields identical results. Otherwise, the consecutive optimization scheme can still be applied.

Relational neural gas shows very robust results in several applications as shown in [14]. Compared to original NG, however, it has the severe drawback that the computation time is $\mathcal{O}(m^2)$, m being the number of data points, and the required space is also quadratic. Thus, this method becomes infeasible for huge data sets. Recently, an intuitive and powerful method has been proposed to extend batch neural gas towards a single pass optimization scheme which can be applied even if the training points do not fit into the main memory [1]. The key idea is to process data in patches, whereby prototypes serve as a sufficient statistics of the already processed data. Here we transfer this idea to relational clustering.

3 The Patch Relational Technique

Assume as before that data are given as a dissimilarity matrix D. During processing of Patch Relational NG, patches of fixed size are cut consecutively from the dissimilarity matrix D, where every patch is a submatrix of D centered around the matrix diagonal.

The idea of the original patch scheme is to add the prototypes from the processing of the former patch P_{i-1} as additional datapoints to the current patch P_i, forming an extended patch P_i^* which includes the previous points in the form of a compressed statistics. The additional datapoints – the former prototypes – are weighted according to the size of their receptive fields, i.e. how many datapoints do they represent in the former patch.

Unlike the situation of original Patch NG [1], where prototypes can simply be converted to datapoints and the inter-patch distances can always be recalculated using the Euclidean metric, the situation becomes more difficult for

relational methods.

In Relational NG prototypes are expressed as convex combinations of unknown Euclidean datapoints, only the distances can be calculated. Moreover, the relational prototypes gained from processing of a patch cannot be simply converted to datapoints for the next patch. They are defined only on the datapoints of the former patch. To calculate the necessary distances between these prototypes and the datapoints of the next patch, the distances between former and next patch must be taken into account, as shown in [14]. But that means touching all elements of the upper half of the distance matrix at least once during processing of all patches, what foils the idea of the patch scheme to reduce computation and memory-access costs.

In this contribution, another way is chosen. In between patches not the relational prototypes itselves but representative datapoints obtained from a so called k-approximation are used to extend the next patch. As for standard patch clustering, the points are equipped with multiplicities. On each extended patch a modified Relational NG is applied taking into account the multiplicities.

4 Summary and Outlook

In the presentation, we presented a special computation scheme based on Relational Neural Gas and Relational Self-Organizing Map that allows to process large dissimilarity datasets, that cannot be hold at once in random-access memory, by a single pass technique of fixed sized patches. The patch size can be chosen to match the given memory constraints. The presented patch version reduces the computation and space complexity with a small loss of accuracy.

In future work, the method will be applied to larger real-world datasets in the context of text processing, musical data mining, and computational biology. Furthermore, a supervision concept as reported in [5] can be integrated. The patch scheme also opens a way towards parallelizing the method as demonstrated in [2].

Algorithm 1: Patch Relational Neural Gas

Begin

Cut the first Patch P_1

Apply Relational NG on $P_1 \longrightarrow$ Relational prototypes W_1

Use k-Approximation on $W_1 \longrightarrow$ Index set N_1

Update Multiplicities m_j according to the receptive fields

Repeat for $t = 2, \ldots, n_p$

Cut patch P_t

Construct Extended Patch P_t^* using P_t and index set N_{t-1}

Apply modified RNG with Multiplicities $\longrightarrow$ Relational prototypes W_t

Use k-Approximation on $W_t \longrightarrow$ Index set N_t

Update Multiplicities m_j according to the receptive fields

Return k-approximation of final prototypes N_{n_p}

End.

References

[1] N. Alex, B. Hammer, and F. Klawonn, Single pass clustering for large data sets, *Proceedings of 6th International Workshop on Self-Organizing Maps* (WSOM 2007), 2007.

[2] N. Alex and B. Hammer, Parallelizing single patch pass clustering, Proc. ESANN 2008

[3] M. Cottrell, B. Hammer, A. Hasenfuss, and T. Villmann (2006), Batch and median neural gas, *Neural Networks*, 19:762-771.

[4] T. Graepel and K. Obermayer (1999), A stochastic self-organizing map for proximity data, *Neural Computation* **11**:139-155.

[5] B. Hammer, A. Hasenfuss, F.-M. Schleif, and T. Villmann (2006), Supervised batch neural gas, In F. Schwenker, and S. Marinai (Eds.), *ANNPR 2006, Springer Lecture Notes in Artificial Intelligence* 4087:33-45.

[6] J.A.Hartigan (1975), Clustering Algorithms, Wiley.

[7] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castren (2003), Trustworthiness and metrics in visualizing similarity of gene expression, *BMC Bioinformatics*, **4:48**.

[8] T. Martinetz, S. Berkovich, and K. Schulten (1993). 'Neural gas' network for vector quantization and its application to time series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569.

[9] T. Villmann, B. Hammer, F. Schleif, T. Geweniger, and W. Herrmann (2006), Fuzzy classification by fuzzy labeled neural gas, *Neural Networks*, **19**:772-779.

[10] Q. Yang and X. Wu (2006), 10 Challenging Problems in Data Mining Research, *International Journal of Information Technology & Decision Making* **5(4)**:597-604.

[11] S. Zhong and J. Ghosh (2003), A unified framework for model-based clustering, *Journal of Machine Learning Research* **4**:1001-1037.

[12] S. Guha, N. Mishra, R. Motwani, L. O'Callaghan (2000). Clustering Data Streams. In *IEEE Symposium on Foundations of Computer Science*, 359-366.

[13] S. Guha, R. Rastogi, K. Shim (1998). CURE: an efficient clustering algorithm for large datasets. In *Proceedings of ACM SIGMOD International Conference on Manag ement of Data*, 73-84.

[14] B. Hammer and A. Hasenfuss, Relational Neural Gas. In J. Hertzberg et al., editors, *KI 2007: Advances in Artificial Intelligence*, Lecture Notes in Artifical Intelligence 4667, pages 190-204, Springer, 2007.

[15] W. Wang, J. Yang, R.R. Muntz (1997). STING: a statistical information grid approach to spatial data mining. In *Proceedings of the 23rd VLDB Conference*, 186-195.

[16] T. Zhang, R. Ramakrishnan, M. Livny (1996). BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the 15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 103-114.

[17] S. Seo and K. Obermayer (2004), Self-organizing maps and clustering methods for matrix data, *Neural Networks* **17:**1211-1230.

[18] T. Villmann, U. Seiffert, F.-M. Schleif, C. Brüß, T. Geweniger and B. Hammer (2006), Fuzzy Labeled Self-Organizing Map with Label-Adjusted Prototypes, In *Proceedings of Conference Artificial Neural Networks in Pattern Recognition (ANNPR) 2006*, F. Schwenker (ed.), Springer, p. 46-56.

[19] T. Kohonen and P. Somervuo (2002), How to make large self-organizing maps for nonvectorial data, *Neural Networks* **15:**945-952.

[20] T. Kohonen (1982), Self-Organized formation of topologically correct feature maps, *Biological Cybernetics*, 43:59-69.

[21] R. Jin, A. Goswami, G. Agrawal (to appear). Fast and Exact Out-of-Core and Distributed K-Means Clustering, *Knowledge and Information System*.

[22] Y. Prudent, A. Ennaji (2005). An incremental growing neural gas learns topology. IJCNN'05.

TU Clausthal
ICOLE 2009, LESSACH, AUSTRIA

Theoretical aspects of kernel GLVQ with differentiable kernel

***Thomas Villmann*[1,2], *Barbara Hammer*[3]**

1 Introduction

Learning vector quantization is mainly influenced by the standard algorithms LVQ1 . . . LVQ3 introduced by KOHONEN [6] as intuitive prototype-based clustering algorithms. These algorithms are heuristically motivated but does not minimize any cost function. Several derivatives were developed to improve the standard algorithms to ensure, for instance, faster convergence, a better adaptation of the receptive fields to optimum Bayesian decision, or an adaptation for complex data structures, to name just a few [5, 7, 9, 12, 15, 4]. GLVQ is an extension of the LVQ2.1 algorithm which avoids the numerical instabilities of LVQ2.1 due to a stochastic gradient descent on a cost function [12]. It has been shown that GLVQ can be seen as a margin optimizer [2].

All the LVQ-algorithms typically are distance based approaches. However, as pointed out in [3] more general *similarity measures* can be considered with the remaining restriction of differentiability. Now the idea is to replace such a general similarity measure by inner products which implies the utilization of *kernels*. In this way we obtain in natural manner a kernel variant of the underlying LVQ algorithms. In particular we will focus on the GLVQ which

[1]E-mail: thomas.villmann@hs-mittweida.de
[2]Department of Mathematics, University of Applied Sciences Mittweida, Mittweida, Germany
[3]Institute of Computer Science, Clausthal University of Technology , Clausthal, Germany

leads to *Kernel-GLVQ*.

2 Inner product based learning vector quantization

To develope the Kernel-GLVQ approach based on differetiable kernels or inner products we shortly review th basics of standard GLVQ. Thereafter the straight forward derivation of Kernel-GLVQ is presented. The section is completed by advisements for relevance learning and hyperparameter adaptation for Kernel-GLVQ.

2.1 Standard GLVQ

Let us first clarify some notations: Let $c_{\mathbf{v}} \in \mathcal{L}$ be the label of input $\mathbf{v}$, $\mathcal{L}$ a set of labels (classes) with $\#\mathcal{L} = N_{\mathcal{L}}$. Let $V \subseteq \mathbb{R}^{D_V}$ be a finite set of inputs $\mathbf{v}$. LVQ uses a fixed number of prototypes (weight vectors, codebook vectors) for each class. Let $\mathbf{W} = \{\mathbf{w_r}\}$ be the set of all codebook vectors and $c_{\mathbf{r}}$ be the class label of $\mathbf{w_r}$. Furthermore, let $\mathbf{W}_c = \{\mathbf{w_r} | c_{\mathbf{r}} = c\}$ be the subset of prototypes assigned to class $c \in \mathcal{L}$. Further, let d be an arbitrary (differentiable) distance measure in V.

We start with the cost function for GLVQ

$$Cost_{GLVQ} = \sum_{\mathbf{v}} \mu(\mathbf{v}) \tag{1}$$

with the *classifier function*

$$\mu(\mathbf{v}) = \frac{d_{\mathbf{r}_+} - d_{\mathbf{r}_-}}{d_{\mathbf{r}_+} + d_{\mathbf{r}_-}} \tag{2}$$

which has to be minimized by gradient descent. Thereby, $d_{\mathbf{r}_+}$ is determined by

$$\mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \text{argmin}_{\mathbf{r} \in A} d(\mathbf{v}, \mathbf{w_r}) \tag{3}$$

and $d_{\mathbf{r}_+} = d(\mathbf{v}, \mathbf{w_s})$ with the additional constraint that $c_{\mathbf{v}} = c_{\mathbf{r}}$, i.e. $d_{\mathbf{r}_+}$ is the squared distance of the input vector $\mathbf{v}$ to the nearest codebook vector labeled with $c_{\mathbf{r}_+} = c_{\mathbf{v}}$. Analogously, $d_{\mathbf{r}_-}$ is defined. Note that $\mu(\mathbf{v})$ is positive if the vector $\mathbf{v}$ is misclassified and negative otherwise.

The learning rule of GRLVQ is obtained taking the derivatives of the above cost function. Using $\frac{\partial \mu(\mathbf{v})}{\partial \mathbf{w}_{\mathbf{r}+}} = \xi^+ \frac{\partial d_{\mathbf{r}+}}{\partial \mathbf{w}_{\mathbf{r}+}}$ and $\frac{\partial \mu(\mathbf{v})}{\partial \mathbf{w}_{\mathbf{r}-}} = \xi^- \frac{\partial d_{\mathbf{r}-}}{\partial \mathbf{w}_{\mathbf{r}-}}$ with

$$\xi^+ = \frac{2 \cdot d_{\mathbf{r}_-}}{(d_{\mathbf{r}_+} + d_{\mathbf{r}_-})^2} \tag{4}$$

and

$$\xi^- = \frac{-2 \cdot d_{\mathbf{r}+}}{(d_{\mathbf{r}_+} + d_{\mathbf{r}_-})^2} \tag{5}$$

one obtains for the weight updates [3]:

$$\triangle \mathbf{w}_{\mathbf{r}_+} = \epsilon^+ \cdot \xi^+ \cdot \frac{\partial d_{\mathbf{r}+}}{\partial \mathbf{w}_{\mathbf{r}+}} \tag{6}$$

$$\triangle \mathbf{w}_{\mathbf{r}_-} = \epsilon^- \cdot \xi^- \cdot \frac{\partial d_{\mathbf{r}-}}{\partial \mathbf{w}_{\mathbf{r}-}} \tag{7}$$

2.2 Inner product based GLVQ and Kernel GLVQ

The idea now is to replace the distance measure in (2) by a *differentiable* (squared) inner product σ. It is obvious that σ defines a norm d_σ. Thus, identifying any subsets by utilization of σ can be done equivalently (in topological sense) by means of the norm d_s and vice versa. In context of GLVQ this implies that all margin analysis is still valid also for inner product based variants of GLVQ. Further, among all inner products σ those are of particular interest, which are generated by kernels κ_ϕ defined in (**??**), i.e. $\sigma = \kappa_\phi$. This motivates the notation *Kernel GLVQ (KGLVQ)*. It should be mentioned here that QIN&SUGANTHAN provided a kernel GRLVQ in [10] for general kernel. However, the derivation of respective prototype updates is complicated. Here, using the differetiability assumption, an alternative much more easy solution is provided. We will outline the theoretical foundations in the following:

In detail, we consider now the inner product σ based classifier function

$$\mu_\sigma(\mathbf{v}) = \frac{\sigma^2_{\mathbf{r}_+} - \sigma^2_{\mathbf{r}_-}}{\sigma^2_{\mathbf{r}_+} + \sigma^2_{\mathbf{r}_-}}$$

which has to be positive iff $\mathbf{v}$ is correctly classified, i.e.

$$\mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \text{argmax}_{\mathbf{r} \in A} \left[(\sigma(\mathbf{v}, \mathbf{w}_\mathbf{r}))^2 \right] \tag{8}$$

and $\sigma_{\mathbf{r}_+}$ as well $\sigma_{\mathbf{r}-}$ play the same role as $d_{\mathbf{r}_+}$ and $d_{\mathbf{r}_-}$. The cost function then reads as

$$Cost_{KGLVQ} = \sum_{\mathbf{v}} \mu_\sigma(\mathbf{v})\,. \tag{9}$$

Using the differentiability assumption we can calculate in complete analogy to the GLVQ above the quantities

$$\frac{\partial \mu_\sigma(\mathbf{v})}{\partial \mathbf{w}_{\mathbf{r}+}} = \xi_\sigma^+ \frac{\partial \sigma_{\mathbf{r}_+}}{\partial \mathbf{w}_{\mathbf{r}+}} \text{ and } \frac{\partial \mu_\sigma(\mathbf{v})}{\partial \mathbf{w}_{\mathbf{r}-}} = \xi_\sigma^- \frac{\partial \sigma_{\mathbf{r}-}}{\partial \mathbf{w}_{\mathbf{r}-}}$$

but now having

$$\xi_\sigma^+ = \frac{4 \cdot \sigma_{\mathbf{r}+} \cdot \sigma_{\mathbf{r}-}^2}{(\sigma_{\mathbf{r}_+}^2 + \sigma_{\mathbf{r}_-}^2)^2} \tag{10}$$

and

$$\xi_\sigma^- = -\frac{4 \cdot \sigma_{\mathbf{r}+}^2 \cdot \sigma_{\mathbf{r}-}}{(\sigma_{\mathbf{r}_+}^2 + \sigma_{\mathbf{r}_-}^2)^2} \tag{11}$$

The final updates for the gradient ascent are immediately obtained as

$$\triangle \mathbf{w}_{\mathbf{r}_+} = \epsilon^+ \cdot \xi_\sigma^+ \cdot \frac{\partial \sigma_{\mathbf{r}_+}}{\partial \mathbf{w}_{\mathbf{r}+}} \tag{12}$$

$$\triangle \mathbf{w}_{\mathbf{r}_-} = \epsilon^- \cdot \xi_\sigma^- \cdot \frac{\partial \sigma_{\mathbf{r}-}}{\partial \mathbf{w}_{\mathbf{r}-}} \tag{13}$$

which contain the derivatives of the inner product or kernel σ.

In case of the usual Euclidean inner product $\sigma_\phi(\mathbf{v}, \mathbf{w}_\mathbf{r}) = \mathbf{v}^T \cdot \mathbf{w}_\mathbf{r}$ with ϕ is the identity function, one simply gets $\frac{\partial \sigma_\phi}{\partial \mathbf{w}_\mathbf{r}} = \mathbf{v}$. Yet, in case of a kernel based inner product κ_ϕ, the derivative of the inner product can easily be carried out without any explicit knowledge of the underlying function ϕ taking into account the kernel trick property. For example, if κ_ϕ is the polynomial kernel $\kappa_\phi = \langle \mathbf{v}, \mathbf{w} \rangle^d$ we have $\frac{\partial \kappa_\phi}{\partial \mathbf{w}} = d \cdot \langle \mathbf{v}, \mathbf{w} \rangle^{d-1} \cdot \mathbf{v}$. For the *rbf-kernel*

$$\kappa_\phi(\mathbf{v}, \mathbf{w}, \gamma) = \exp\left(-\frac{(\mathbf{v} - \mathbf{w})^2}{2\gamma^2}\right) \tag{14}$$

one obtains $\frac{\partial \kappa_\phi}{\partial \mathbf{w}} = \frac{1}{\gamma^2} \exp\left(-\frac{(\mathbf{v}-\mathbf{w})^2}{2\gamma^2}\right)(\mathbf{v} - \mathbf{w})$ whereas for the exponential kernel $\kappa_\phi = \exp(\langle \mathbf{v}, \mathbf{w} \rangle)$ this procedure yields $\frac{\partial \kappa_\phi}{\partial \mathbf{w}} = \exp(\langle \mathbf{v}, \mathbf{w} \rangle) \cdot \mathbf{v}$. For further kernel examples we refer to [14], Chapt. 9.

Another widely applied inner product is the *Sobolev inner product of degree* K

$$\begin{aligned} \langle f,g\rangle_{\mathcal{S},K} &= \langle f,g\rangle_{\mathrm{E}} + \sum_{1\leq j\leq K} \left\langle D^{(j)}f, D^{(j)}g \right\rangle_{\mathrm{E}} && (15)\\ &\asymp \langle f,g\rangle_{\mathrm{E}} + \left\langle D^{(K)}f, D^{(K)}g \right\rangle_{\mathrm{E}} && (16) \end{aligned}$$

with differential operators $D^{(k)}$ defining the k-th derivative and f, g are functions or vectorial representations thereof [8]. Sobolev norms and inner products play a fundamental role in functional data analysis, such as for time series analysis, sequence processing or spectral data analysis in biomedicine, geo- and astrophysics, chemistry etc. [11],[16],[17]. Using $\langle f,g\rangle_{\mathcal{S},K}$ in KGLVQ results an appropriate inner products based scheme for functional data classification.

2.3 Parameter adaptation for Gaussian kernels

Kernel width

The width γ of the Gaussian kernel (14) crucially influences the performance of the classifier. Hence, a careful adjustment is mandatory. Yet, an alternative is to individualize the kernel width $\gamma_{\mathbf{r}}$ for each prototype $\mathbf{w_r}$ and, afterwards one keeps them as parameters to be learned, too [13]. As the prototypes itself, this should be done by stochastic gradient ascent on $Cost_{KGLVQ}$, i.e. we consider $\frac{\partial Cost_{KGLVQ}}{\partial \gamma_{\mathbf{r}}}$. In particular, we have to calculate

$$\begin{aligned} \frac{\partial \mu_\sigma(\mathbf{v})}{\partial \gamma_{\mathbf{r}_+}} &= \frac{\partial}{\partial \gamma_{\mathbf{r}_+}} \left[\frac{\sigma^2_{\mathbf{r}_+} - \sigma^2_{\mathbf{r}_-}}{\sigma^2_{\mathbf{r}_+} + \sigma^2_{\mathbf{r}_-}} \right] \\ &= \xi^+_\sigma \cdot \frac{\partial \sigma_{\mathbf{r}_+}}{\partial \gamma_{\mathbf{r}_+}} \end{aligned}$$

determining the adaptation $\triangle\gamma_{\mathbf{r}+}$, which yields for the localized Gaussian kernel (14)

$$\begin{aligned} \frac{\partial \mu_\sigma(\mathbf{v})}{\partial \gamma_{\mathbf{r}_+}} &= \xi^+_\sigma \cdot \frac{\partial \kappa_\phi\left(\mathbf{v}, \mathbf{w}_{\mathbf{r}_+}, \gamma_{\mathbf{r}_+}\right)}{\partial \gamma_{\mathbf{r}_+}} \\ &= \xi^+_\sigma \cdot \frac{\kappa_\phi\left(\mathbf{v}, \mathbf{w}_{\mathbf{r}_+}, \gamma_{\mathbf{r}_+}\right)}{\gamma^3_{\mathbf{r}_+}} \cdot \left(\mathbf{v} - \mathbf{w}_{\mathbf{r}_+}\right)^2 . \end{aligned}$$

Analogously, we find

$$\frac{\partial \mu_\sigma(\mathbf{v})}{\partial \gamma_{\mathbf{r}_-}} = \xi_\sigma^- \cdot \frac{\kappa_\phi\left(\mathbf{v}, \mathbf{w}_{\mathbf{r}_-}, \gamma_{\mathbf{r}_-}\right)}{\gamma_{\mathbf{r}-}^3} \cdot \left(\mathbf{v} - \mathbf{w}_{\mathbf{r}_-}\right)^2 .$$

Relevance learning

The Gaussian kernel takes as ingredients usually the Euclidean norm of the vector difference. However, other norms are also possible, for example one could think about special choices for specific problems as Sobolev-norms for functional data analysis. Here we draw the attention to the so-called scaled Euclidean metric

$$d^\lambda(\mathbf{v}, \mathbf{w}) = \sum_i \lambda_i \cdot (v_i - w_i)^2$$

with $\sum_i \lambda_i = 1$. As usual in relevance learning [3], the scaling parameters λ_i can be adapted with respect to the classification task at hand by gradient learning, which leads again to a gradient ascent but now as $\frac{\partial Cost_{KGLVQ}}{\partial \lambda_i}$. Here we have to consider $\frac{\partial \mu_\sigma(\mathbf{v})}{\partial \lambda_i}$. We obtain for $\mathbf{w}_{\mathbf{r}\pm}$

$$\begin{aligned}
\frac{\partial \mu_\sigma(\mathbf{v})}{\partial \lambda_i} &= \xi_\sigma^+ \cdot \frac{\partial \kappa_\phi\left(\mathbf{v}, \mathbf{w}_{\mathbf{r}_\pm}, \gamma_{\mathbf{r}_\pm}\right)}{\partial \lambda_i} \\
&= -\xi_\sigma^+ \cdot \frac{\kappa_\phi\left(\mathbf{v}, \mathbf{w}_{\mathbf{r}_\pm}, \gamma_{\mathbf{r}_\pm}\right)}{2\gamma^2} \left(v_{\mathbf{r}_\pm, i} - w_{\mathbf{r}_\pm, i}\right)^2
\end{aligned}$$

to be plugged into a respective gradient learning as usual. We denote this approach as *Kernelized Relevance GLVQ (KGRLVQ)*.

Clearly, one could think about more sophisticated relevance learning schemes like *matrix learning*, where the metric is given by a general bilinear form

$$\begin{aligned}
d^{\mathbf{M}}(\mathbf{v}, \mathbf{w}) &= (\mathbf{v} - \mathbf{w})^T \cdot \mathbf{M} \cdot (\mathbf{v} - \mathbf{w}) \\
&= (\mathbf{v} - \mathbf{w})^T \cdot \mathbf{\Omega}^T \mathbf{\Omega} \cdot (\mathbf{v} - \mathbf{w})
\end{aligned}$$

with $\mathbf{M}$ being a (symmetric) positive definite matrix [1]. The derivation is straight forward containing the derivativs

$$\frac{\partial d^{\mathbf{M}}(\mathbf{v}, \mathbf{w})}{\partial \Omega_{jm}} = \sum_j (v_k - w_k)\Omega_{jm}(v_j - w_j) + \sum_i (v_i - w_i)\Omega_{im}(v_k - w_k)$$

or, in matrix notation,

$$\frac{\partial d^{\mathbf{M}}(\mathbf{v},\mathbf{w})}{\partial \mathbf{\Omega}} = 2(\mathbf{v}-\mathbf{w})\left[\mathbf{\Omega}(\mathbf{v}-\mathbf{w})\right]^T$$

and yields *Kernelized Matrix GLVQ (KGMLVQ)*.

3 Conclusion

In the present article we give the theoretical foundations for Kernel-GRLVQ with differentiable kernels. This approach is an easy alternative to the earlier developed kernel variant of GLVQ which does not assume the differentiability of the used kernel. The achieved update scheme for the prototypes is similar to the original GLVQ. Further, we give perspectives to integrate relevance learning and hyperparameter optimization into Kernel-GRLVQ for classification task dependent parameter estimation for optimum classification.

References

[1] M. Biehl, B. Hammer, F.-M. Schleif, P. Schneider, and T. . Villmann. Stationarity of matrix relevance learning vector quantization. *Machine Learning Reports*, 3(MLR-01-2007):1–17, 2009. ISSN:1865-3960, http://www.uni-leipzig.de/~compint/mlr/mlr-01-2009.pdf.

[2] K. Crammer, R. Gilad-Bachrach, A.Navot, and A.Tishby. Margin analysis of the LVQ algorithm. In *Proc. NIPS 2002*, http://www-2.cs.cmu.edu/Groups/NIPS/NIPS2002/NIPS2002preproceedings/index.html, 2002.

[3] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[4] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15(7):1589–1604, 2003.

[5] B. Hammer and T. Villmann. Mathematical aspects of neural networks. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2003)*, pages 59–72, Brussels, Belgium, 2003. d-side.

[6] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[7] T. Kohonen, S. Kaski, H. Lappalainen, and J. Saljärvi. The adaptive-subspace self–organizing map (assom). In *International Workshop on Self–Organizing Maps (WSOM'97)*, pages 191–196, Helsinki, 1997.

[8] A. Kolmogorov and S. Fomin. *Reelle Funktionen und Funktionalanalysis*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1975.

[9] M. Pregenzer, G. Pfurtscheller, and D. Flotzinger. Automated feature selection with a distinction sensitive learning vector quantizer. *Neurocomputing*, 11(1):19–29, 1996.

[10] A. Qin and P. Suganthan. A novel kernel prototype-based learning algorithm. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 4, pages 621–624, 2004.

[11] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Science+Media, New York, 2nd edition, 2006.

[12] A. S. Sato and K. Yamada. Generalized learning vector quantization. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 423–429. MIT Press, 1995.

[13] P. Schneider and M. Biehl and B. Hammer. Hyperparameter Learning in Robust Soft LVQ. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2009)*, pages 517–552, Brussels, Belgium, 2009. d-side.

[14] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis and Discovery*. Cambridge University Press, 2004.

[15] P. Somervuo and T. Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10(2):151–159, 1999.

[16] T. Villmann and B. Hammer. Functional principal component learning using oja's method and sobolev norms. In J. Principe, editor, *Advances in Self-Organizing Maps - Proceeding of the Workshop on Self-Oranizing Maps (WSOM)*, pages 325–333. Springer, 2009.

[17] T. Villmann and F.-M. Schleif. Functional vector quantization by neural maps. In *Proceedings of WHISPERS 2009*, page in press, 2009.